

**Sex and Parental Genome Effects on
Bovine
Fetal Development**

Ruijie Liu

A thesis submitted for the degree of

Doctor of Philosophy

School of Animal and Veterinary Sciences

Faculty of Sciences

The University of Adelaide

August 2020

Contents

Declaration	x
Publication/prepared manuscript	xi
Abstract	xix
Acknowledgements	xxi
1 Literature Review	1
1.1 Mammalian genome	1
1.1.1 Genome components	1
1.1.2 Gene expression and regulation	3
1.1.3 Epigenetics	4
1.1.3.1 Dosage compensation	4
1.1.3.2 Genes that escape X inactivation	5
1.1.3.3 Genomic imprinting	6
1.2 Mammalian development	6
1.2.1 Mammalian prenatal development	6
1.2.2 Sexual dimorphism during development	8
1.3 Next Generation Sequencing technology	9
1.3.1 Short and long read sequencing	9
1.3.2 Transcriptome sequencing	10
1.4 Bioinformatics methods	12
1.4.1 Steps of genome assembly	12
1.4.1.1 <i>De novo</i> sequence assembly using long read data	12
1.4.1.2 Scaffolding	13
1.4.2 Bioinformatic processes in transcriptome analysis	14
1.4.2.1 Quality control and alignment	14
1.4.2.2 Variant Calling	15
1.4.2.3 Differential gene expression	15

1.4.2.4	Pathway analysis	16
1.5	Introduction to cattle	17
1.5.1	Cattle domestication	17
1.5.2	<i>Bos taurus taurus</i> and <i>Bos taurus indicus</i>	17
1.5.3	Genetic variation	18
1.5.4	Cattle genome assembly	20
1.6	Research aim	22
	References	24
2	New Insights into Mammalian Sex Chromosome Structure and Evolution using High-Quality Sequences from Bovine X and Y Chromosomes	44
2.1	Supplementary Notes	56
2.1.1	X chromosome scaffolds identification and orientation	56
2.1.2	Y chromosome scaffolds identification and orientation	56
2.1.3	Comparison of X and Y chromosomes in mammals	57
2.1.4	Gene annotation of sex chromosomes	57
2.2	Supplementary Figures	59
2.3	Supplementary Tables	63
	References	67
3	X-Y chromosome gametologues explain sex differences in fetal organ weights	69
3.1	Abstract	70
3.2	Main	70
	References	75
3.3	Methods	78
3.3.1	Animals, phenotypes and tissue sampling	78
3.3.2	RNA preparation and sequencing	78
3.3.3	Sex-specific expression analysis	78
3.3.4	Dosage compensation analysis	79
3.3.5	Gametologue expression analysis	80
3.3.6	Sex variation in organ weight explained by gametologues	80

References	82
3.4 Supplementary Figures	84
3.5 Supplementary Tables	93
4 Different cattle breeds show distinctive gene expression patterns, including imprinting signatures in reciprocal crosses	101
4.1 Abstract	102
4.2 Introduction	102
4.3 Material and Methods	104
4.3.1 Animals and sample collection	104
4.3.2 RNA isolation, library preparation and sequencing	104
4.3.3 Data analysis	104
4.3.4 Functional analysis of DEGs	105
4.3.5 Identification of Brahman/Angus gene expression pattern in crossbred groups	106
4.4 Results	107
4.4.1 Expression profile of five tissues	107
4.4.2 Differential gene expression between purebred groups	107
4.4.3 DE genes common to all five tissues	108
4.4.4 Tissue-specific genes between purebred groups	109
4.4.5 Differential gene expression between crossbred groups	110
4.4.6 Expression pattern of DEGs from the purebred cattle in comparison with crossbred groups	111
4.5 Discussion	112
References	117
4.6 Supplementary Figures	124
4.7 Supplementary Tables	126
5 General Discussion	131
References	138

6	Supporting Publication: Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle	142
6.1	Supplementary Figures	157
6.2	Supplementary Tables	166
6.3	Supplementary Notes	178
6.3.1	Comparison of different Hi-C scaffolding programs	178
6.3.2	Comparison of optical map based scaffolding approaches	179
6.3.2.1	<i>De novo</i> optical map assembly and haplotype resolution	179
6.3.2.2	Haplotype-resolved scaffold assembly	180
6.3.2.3	Conventional optical map scaffold assembly	181
6.3.3	Genome annotation of UOA_Brahman_1 using the NCBI annotation pipeline	182
6.3.4	Genome annotation of UOA_Angus_1 and UOA_Brahman_1 using the Ensembl annotation pipeline	183
6.3.5	Further assembly evaluation	185
6.3.6	Identification of selective sweep regions	186
6.3.6.1	The selection sweep method	186
6.3.6.2	Pathway analysis	189
6.3.6.3	Interpretation of selective sweep results	189
6.3.7	Further Iso-Seq analysis	191
6.3.7.1	SNP calls missed by IsoPhase are either in homopolymer regions or have low coverage	191
6.3.7.2	IsoPhase-unique SNP calls are dominantly A to G calls, which suggests RNA editing	192
6.3.8	Rarefaction analysis of covered genes and transcripts	193
	References	194

List of Figures

1.1	Chromatin structure	2
1.2	Steps in gene expression	3
1.3	Neighbour-net analysis for 174 populations of taurine and indicine cattle .	19
2.1	Alignment of the Brahman X with other mammalian X chromosomes. . .	59
2.1	(continued) Alignment of the Brahman X with other mammalian X chromosomes.	60
2.2	Alignment of the Angus cattle Y against other mammalian Y chromosomes.	61
2.3	Multi-copy genes in Angus Y ampliconic region.	62
3.1	Differentially expressed genes on the fetal sex chromosomes discriminating females and males at midgestation (Day153).	71
3.2	Female-male differences in gametologue expression explain sex effects on fetal organ weights at midgestation (Day153).	73
3.3	Multi-dimensional scaling plots of male-female difference on a two-dimensional scatterplot.	84
3.4	Venn diagram with numbers of differentially expressed genes (FDR <0.05) between males and females in five fetal tissues.	85
3.5	Venn diagram showing the overlap between differentially expressed (DE) genes (FDR<0.05) between males and females in fetal and adult tissues. .	86
3.6	Ratio of the median expression levels of X-specific genes and autosomal genes of female (red) and male (blue) samples.	87
3.7	Heatmap of differentially expressed (DE) genes between females and males and ratios of the median expression levels of X chromosome genes and autosomal genes in female and male fetal tissues.	88
3.8	Comparison of differentially expressed (DE) genes between females and males in fetal and adult tissues.	89
3.9	Comparison of mean XX:X, X:Y and XX:combined XY expression ratios of gametologues in two fetal (F) and adult (A) tissues.	90
3.10	Organ weights of females and males and expression levels of gametologues that explain variation captured by the factor 'sex' in linear models.	91

3.11	Correlations between expression levels of gametologues within tissues.	92
4.1	Multi-dimensional scaling (MDS) plot of sample expression profiles in five tissues.	108
4.2	Venn diagram with numbers of differentially expressed genes across five tissues and their pathways	109
4.3	Examples of expression patterns among genotype groups.	111
4.4	Comparison of gene expression levels in five tissues for pure Angus and Brahman.	124
4.5	Multi-dimensional scaling plots reveals genetic group difference in gene expression profiles in each tissue.	125
6.1	Comparison of chromosome sizes between Angus, Brahman and Hereford assemblies.	157
6.2	Distribution of un-gapped contig lengths in the three cattle breeds (UOA_Angus_1, UOA_Brahman_1 and ARS-UCD1.2) and water buffalo (UOA_WB_1).	158
6.3	The count in (\log_{10} scale) of LINE/L1, LINE/RTE-BovB and Satellite/centromeric repeats in cattle genome assemblies.	159
6.4	Coverage plot of <i>FADS2P1</i> in individuals of different cattle breeds.	160
6.5	Distribution and breed specificity of Brahman and Angus structural variants.	161
6.6	Analysis of copy number variations using different reference assemblies.	162
6.7	Full-length Iso-Seq transcripts (bottom) and RNA-Seq coverage for <i>ARIH2</i> in the brain tissue (top).	163
6.8	Normalized tissue-specific transcript counts for genes with allelic imbalance and higher expression of the Brahman allele in brain.	164
6.9	Distribution of unassigned PacBio WGS read length.	165
6.10	Histograms of the mean proportion alternate allele in Brahman and the other six taurine breeds.	188
6.11	Selective sweep analysis in Brahman.	190
6.12	Percentages of SNP type across those unique in each of PacBio Iso-Seq, RNA-Seq and genome WGS datasets.	192

6.13 Rarefaction analysis of seven tissues in F1 animal at the level of a) gene, b) transcript.	193
--	-----

List of Tables

2.1	Summary of X and Y chromosome protein-coding genes	63
2.2	Summary of X and Y chromosome protein-coding genes	64
2.3	Summary of X and Y chromosome protein-coding genes	65
2.4	Copy numbers of OBP and BDA20 genes in each species	66
3.1	Number of expressed genes by chromosome and fetal tissue. Y* chromosome here refers to only the non-PAR Y sequence.	93
3.2	Nucleotide and protein sequence identity of X- and Y-chromosome gametologues.	94
3.3	Differentially expressed genes between females and males in five tissues (FDR <0.05).	95
3.4	Summary statistics of phenotype data of fetuses and of expression values for XX and combined XY gametologues.	97
3.5	Mean organ weights and variation explained by the sex effect.	98
3.6	SAS Proc GLMSelect results for gametologue subset selection by organ.	99
3.7	Proportion of Sex Effect Variation in Organ Weight Explained by Gametologue Subsets.	100
4.1	Number of genes showing a parent of origin effect on expression patterns in five tissues.	112
4.2	Highly expressed genes (average CPM) in five tissues.	126
4.3	Significant tissue specific gene ontology pathways.	129
6.1	Annotation features in Brahman, Angus and Hereford assemblies.	166
6.2	Assembly quality score values.	167
6.3	BUSCO assessment of the completeness of single-copy orthologs for Angus and Brahman genomes.	168
6.4	Site models of CODEML for FADS2P1 and positively selected sites.	169
6.5	Genes identified in the selective sweep intervals.	170
6.6	Annotation of SNP and INDEL variants.	174
6.7	Breed-specific structural variant (SV) type and over/under- represented gene ontology for biological processes.	175

6.8	Input dataset used to perform optical map-based assemblies using the haplotype-resolved versus the conventional haplotype unaware approach.	180
6.9	Angus-selected Bionano assembly with Angus contigs.	181
6.10	Brahman-selected Bionano assembly with Brahman contigs.	181
6.11	Offspring scaffolding with Angus contigs.	182
6.12	Offspring scaffolding with Brahman contigs.	182
6.13	Comparisons of the number of corrected coding sequences in selected mammalian species.	183
6.14	Initial transcript models for each major input data type for Ensembl annotation.	184
6.15	Comparisons of different window sizes for SNP counts and consecutive SNP distances in each window.	188

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

August 2020

Ruijie Liu

Publication/prepared manuscript

Chapter 2: New Insights into Mammalian Sex Chromosome Structure and Evolution using High-Quality Sequences from Bovine X and Y Chromosomes

Ruijie Liu¹, Wai Yee Low¹, Rick Tearle¹, Sergey Koren², Jay Ghurye³, Arang Rhie², Adam M. Phillippy², Benjamin D. Rosen⁴, Derek M. Bickhart⁵, Timothy P.L. Smith⁶, Stefan Hiendleder¹, John L. Williams¹

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia

²Genomic Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

³Center for Bioinformatics and Computational Biology, Lab 3104A, Biomolecular Science Building, University of Maryland, College Park, Maryland, USA

⁴Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, Maryland, USA

⁵Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, Wisconsin, USA

⁶US Meat Animal Research Centre, ARS USDA, Clay Centre, Nebraska, USA

Published in 2019, BMC Genomics 20 (1), 1-11

Statement of Authorship

Title of Paper	New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Liu, R., Low, W.Y., Tearle, R., Koren, S., Ghurye, J., Rhie, A., Phillippy, A.M., Rosen, B.D., Bickhart, D.M., Smith, T.P., Hiendleder, S. and Williams J. L., 2019. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. BMC genomics, 20(1), pp.1-11.

Principal Author

Name of Principal Author (Candidate)	Ruijie Liu			
Contribution to the Paper	validated the sex chromosome assemblies using various bovine markers; annotated the sex chromosomes and compared them to other mammals; drafted the manuscript			
Overall percentage (%)	60%			
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 80%;"></td> <td style="width: 20%;">Date</td> <td>25 July 2020</td> </tr> </table>		Date	25 July 2020
	Date	25 July 2020		

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Wai Yee Low			
Contribution to the Paper	consolidated all data on assemblies to produce final chromosome-level haplotype-resolved genomes; drafted the manuscript			
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 80%;"></td> <td style="width: 20%;">Date</td> <td>16 July 2020</td> </tr> </table>		Date	16 July 2020
	Date	16 July 2020		

Name of Co-Author	Rick Tearle		
Contribution to the Paper	drafted the manuscript		
Signature		Date	July 16 2020

Name of Co-Author	Sergey Koren		
Contribution to the Paper	assembled the contigs and provided guidance on scaffolding; provided the Hi-C scaffolds.		
Signature		Date	13/04/2020

Name of Co-Author	Jay Ghurye		
Contribution to the Paper	provided the Hi-C scaffolds.		
Signature		Date	13/04/2020

Name of Co-Author	Arang Rhie		
Contribution to the Paper	assembled the contigs and provided guidance on scaffolding.		
Signature		Date	Jul. 1. 2020

Name of Co-Author	Adam M. Phillippy		
Contribution to the Paper	assembled the contigs and provided guidance on scaffolding.		
Signature		Date	April 13, 2020

Name of Co-Author	Benjamin D. Rosen		
Contribution to the Paper	compared scaffolding using various programs; validated the sex chromosome assemblies using various bovine markers.		
Signature		Date	13/04/2020

Name of Co-Author	Derek M. Birckhart		
Contribution to the Paper	compared scaffolding using various programs.		
Signature		Date	4/13/2020

Name of Co-Author	Timothy P.L. Smith		
Contribution to the Paper	conceived and managed the project; created, sequenced, and curated sequencing data.		
Signature		Date	April 17, 2020

Name of Co-Author	Stefan Hiendleder		
Contribution to the Paper	designed breeding experiments, provided dam mtDNA typing, tissue and DNA samples and clarified the single PAR in cattle using BAC information; drafted the manuscript		
Signature		Date	16-07-2020

Name of Co-Author	John L. Williams		
Contribution to the Paper	conceived and managed the project; drafted the manuscript		
Signature		Date	19 th July 2020

Chapter 3: X-Y chromosome gametologues explain sex differences in fetal organ weights

Ruijie Liu¹, Rick Tearle¹, Wai Yee Low¹, Tong Chen¹, Dana Thomsen^{1,2}, Consuelo Amor S. Estrella^{1,2,4}, Ruidong Xiang^{1,2,3}, David R. Rutley¹, Timothy P.L. Smith⁵, John L. Williams¹, Stefan Hiendleder^{1,2}

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia

²Robinson Research Institute, The University of Adelaide, Adelaide, Australia

³Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, Australia

⁴University of the Philippines, Laguna, Philippines

⁵USMARC, USDA-ARS-US Meat Animal Research Center, Clay Center, NE, USA

Intended for submission to Nature Genetics in Brief Communication format

Statement of Authorship

Title of Paper	X-Y chromosome gametologues explain sex differences in fetal organ weights
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Intended to submit to Nature Genetics in Brief Communication format Ruijie Liu, Rick Tearle, Wai Yee Low, Tong Chen, Dana A. Thomsen, Consuelo Amor S. Estrella, Ruidong Xiang, David L. Rutley, Timothy P.L. Smith, John L. Williams, Stefan Hiendleder (2020). X-Y chromosome gametologues explain sex differences in fetal organ weights

Principal Author

Name of Principal Author (Candidate)	Ruijie Liu
Contribution to the Paper	Analysed and interpreted data, drafted the manuscript
Overall percentage (%)	60%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	_____
Date	July 25 2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Rick Tearle
Contribution to the Paper	interpreted data
Signature	_____
Date	July 16 2020

Name of Co-Author	Wai Yee Low		
Contribution to the Paper	interpreted data		
Signature		Date	July 16 2020

Name of Co-Author	Tong Chen		
Contribution to the Paper	managed sample and phenotype resources, extracted and performed QC of RNA samples		
Signature		Date	14/07/2020

Name of Co-Author	Dana A. Thomsen		
Contribution to the Paper	managed sample and phenotype resources, extracted and performed QC of RNA samples		
Signature		Date	25 July 2020

Name of Co-Author	Consuelo Amor S. Estrella		
Contribution to the Paper	collected and analysed phenotypes		
Signature		Date	19 July 2020

Name of Co-Author	Ruidong Xiang		
Contribution to the Paper	collected and analysed phenotypes		
Signature		Date	19/07/2020

Name of Co-Author	David L. Rutley		
Contribution to the Paper	analysed and interpreted data		
Signature		Date	27/7/20

Name of Co-Author	Timothy P.L. Smith		
Contribution to the Paper	conceived and managed the project; generated RNA-Seq data		
Signature		Date	14 July 2020

Name of Co-Author	John L. Williams		
Contribution to the Paper	conceived and managed the project; generated RNA-Seq data; interpreted data		
Signature		Date	19 th July 2020

Name of Co-Author	Stefan Hiendleder		
Contribution to the Paper	conceived and managed the project; designed and managed animal experiments, sampling; analysed and interpreted data, drafted the manuscript		
Signature		Date	16-07-2020

Abstract

During fetal development, the process of forming organs and tissues is mediated by tissue-specific patterns of gene expression. Studying qualitative and quantitative changes in the transcriptome and understanding the mechanisms that regulate gene expression and the association with specific phenotypes in bovine fetal development will help us to explore the sex effect and breed effect. To carry out this work, a well-assembled cattle reference genome is essential, but the current cattle reference genome is incomplete and in particular, missing the Y chromosome.

In this thesis I describe the first bovine sex chromosome assemblies for *Bos taurus indicus* and *Bos taurus taurus* cattle, that include the complete pseudoautosomal regions (PAR), which span 6.84 Mb and comprises 31 genes, and three Y chromosome X-degenerate (X-d) regions. The results show the ruminant PAR boundary is at a similar position to those of the pig and dog, but that the ruminant PAR extends further than those of human and horse. Differences in the PAR boundaries are consistent with evolutionary divergence times. A bovidae-specific expansion of members of the lipocalin gene family in the PAR reported here, may affect immune-modulation and anti-inflammatory responses in ruminants. Comparison of the X-d regions of Y chromosomes across species revealed that five of the X-Y gametologues, which are known to be global regulators of gene activity and candidate sexual dimorphism genes, are conserved.

I report the transcriptome sequencing of 120 samples (60 males and 60 females) and analyzed differences in gene expression between male and female tissues derived from all three germ layers of the embryo, including brain, liver and lung, skeletal muscle and placenta. A remarkably small set of XY genes (gametologues) was identified that differentiate males and females across all tissues. Expression levels of paired gametologues in males and females are unbalanced and explain 18% - 96% of the phenotypic variance in organ weights attributed to the sex effect. Considering the significant programming events at the embryo-fetal stage, we propose that early differences in gametologue expression between females and males are fundamental drivers of phenotypic differences between the sexes.

The 120 samples used in this study were from 4 genetic groups: pure Angus, pure Brahman and their reciprocal crosses. Differential gene expression between the pure breed individuals and between the reciprocal crosses was explored. There were 110 genes differentially expressed (DEGs) between pure Angus and pure Brahman in all tissues which were related to functions including immune response and stress response. The DEG between the purebred groups and in the reciprocal crosses showed an additive expression pattern, where both paternal and maternal genomes contributed to the gene expression levels. Only 5% of DEGs in each tissue showed a parent of origin driven expression, Angus or Brahman, and showed both maternal and paternal dominant effects.

In summary, the newly assembled cattle sex chromosomes helped us to identify the PAR, X-degenerate region and the locations of gametologues which provide a clear reference for sex-specific study. Studies of sex-specific and breed-specific effects on fetal development showed gametologues play a major role in early female-male phenotypic differentiation which also provided solid evidence to support further parent of origin studies.

Acknowledgements

Since I commenced this project 3.5 years ago, many people have provided significant help in this project and it is not possible I could have done everything alone. Here I would like to express my sincere thanks to every contributor:

Great thanks to my principal supervisor John Williams for giving me a chance to undertake this project; for providing an opportunity to travel to the US to work with our world-best collaborators; for supporting me to attend many national and international conferences; for comforting and encouraging me when I was too stressful, and of course, for bearing my poor English writing and biological knowledge. I learned a lot from him, not just valuable for my PhD studies, but also for my whole academic career. It is my honour to be his last PhD student.

I would like to thank my co-supervisor Stefan Hiendleder, for his great help, guidance and support for all my projects. He taught me a lot about biology which makes me feel confident to communicate with any biologist.

I would like to express my thanks to my co-supervisor David Adelson for his support in sex chromosome assembly project.

A significant thanks must go to Rick Tearle. Although he is not one of my supervisors, he still spent a lot of his time to help me through the last 3.5 years, giving me guidance in various of bioinformatics topics, and most importantly, helped me to finalise my thesis.

I would like to thank Lloyd Low for his guidance and help in sex chromosome assembly project and thank Cindy Bottema for patiently giving me suggestions and feedback in all my projects!

My sincere thanks are also extended to Australia Research Training Program and J.S. Davies Research Centre for providing the PhD scholarship.

In the end, I own huge gratitude to my parents, my husband Lei and my baby girl Cheryl. To my parents, thank you for supporting me to study in Australia in the last 15 years, from bachelor's degree to PhD and also agreeing with every decision I have made, no matter if

you liked it or not. To Lei, thank you for supporting every decision in my career, thank you for your sacrifice in many ways, I really appreciate your help and contribution to our family and also bearing my bad temper! To my newborn Cheryl, you have been quietly accompanying me in my last 10 months writing time, you really accelerate my progress of PhD thesis writing!

1 Literature Review

1.1 Mammalian genome

Deoxyribonucleic acid (DNA) is composed of four different bases: adenine (A), guanine (G), cytosine (C) and thymine (T) (Alberts et al., 2002). DNA is the biological molecule which contains heritable information that plays a central role in the development and maintenance of an organism (Watson and Crick, 1953; Alberts et al., 2002). The ~3 billion base pairs (3Gb) of a typical mammalian genome is 2-3 metres long and is compacted into the cell nucleus which is only 5-10 μm in diameter. This is achieved by wrapping the DNA around eight highly alkaline proteins called histones. There are four highly conserved core histones: H2A, H2B, H3 and H4 (Alberts et al., 2002). This DNA-histone structure is a nucleosome which is also known as the basic unit of chromatin (Figure 1.1). Chromatin is further condensed into individual chromosomes when cells divide. For example, the human genome is packaged in 22 pairs of autosomes and 1 pair of sex chromosomes (Alberts et al., 2002) plus a mitochondrial genome. The human genome contains ~21,300 coding genes (Pertea et al., 2018) which are the segments of DNA with instructions for proteins, which form the cell structure and dictate cell function. The DNA is transcribed into ribonucleic acid (RNA) first, and then translated into protein.

1.1.1 Genome components

The mammalian genome includes both protein-coding genes and non-coding DNA (as well as mitochondrial DNA). Protein coding genes are interleaved with exons (coding sequence) and introns (non-coding sequence). Primary RNA transcripts contain both, with introns spliced out during RNA processing, leaving only exons in a mature messenger RNA (mRNA) transcript, to which a 5' cap and 3' poly-a tail are added (Licatalosi and Darnell, 2010). During translation, mRNA is associated with a ribosome in the cytoplasm (Levine and Tjian, 2003). Transfer RNAs (tRNAs) with their attached amino acids bind to mRNA by matching codon sequences, and the amino acids are assembled into proteins (Figure 1.2). Protein coding genes make up ~50% of the mammalian genomic DNA but only ~2% of the genome is in exons (Lander et al., 2001). Repetitive DNA and non-coding sequences

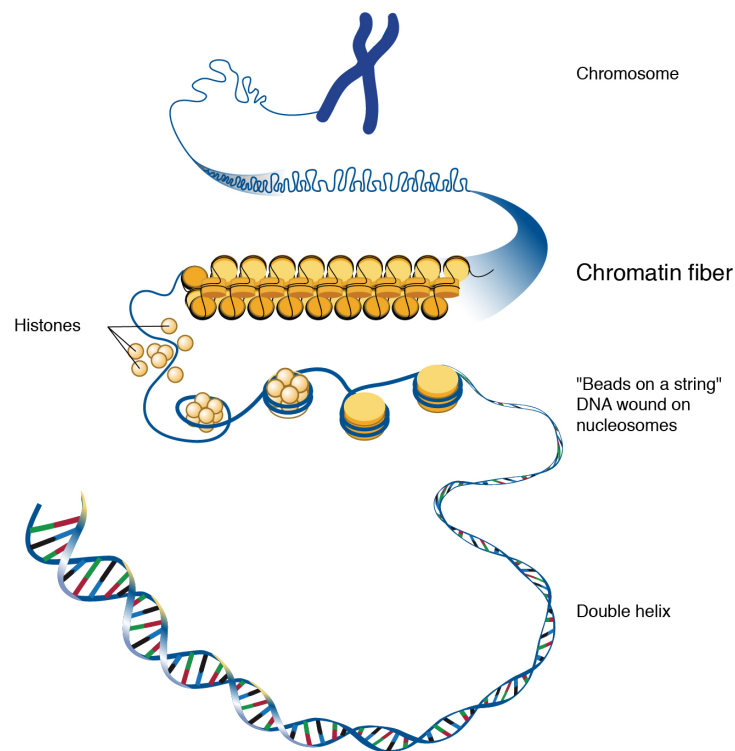


Figure 1.1: Chromatin structure (National human genome research institute (NHGRI, 2020)).

contribute the remaining ~50%. Two classes of non-coding RNA encoded by the genome are microRNAs (miRNAs), which are short RNA molecules that regulate translation and protein degradation (Oliveto et al., 2017), and long non-coding RNAs (lncRNAs), that play many important roles in regulation of gene expression (Fernandes et al., 2019).

Because non-coding sequences do not code for proteins, for a long time non-coding DNA was considered as “junk DNA” with no known purpose (Palazzo and Lee, 2015). However, as the biology developed, some of non-coding DNA was found to be integral in gene regulation (Fang et al., 2019). For example, non-coding DNA could regulate elements and determine when and where genes are turned on and off (Todeschini et al., 2014). Non-coding regions include promoter sequences that provide binding sites for the protein machinery that help activate transcription (Shlyueva et al., 2014; Haberle and Stark, 2018).

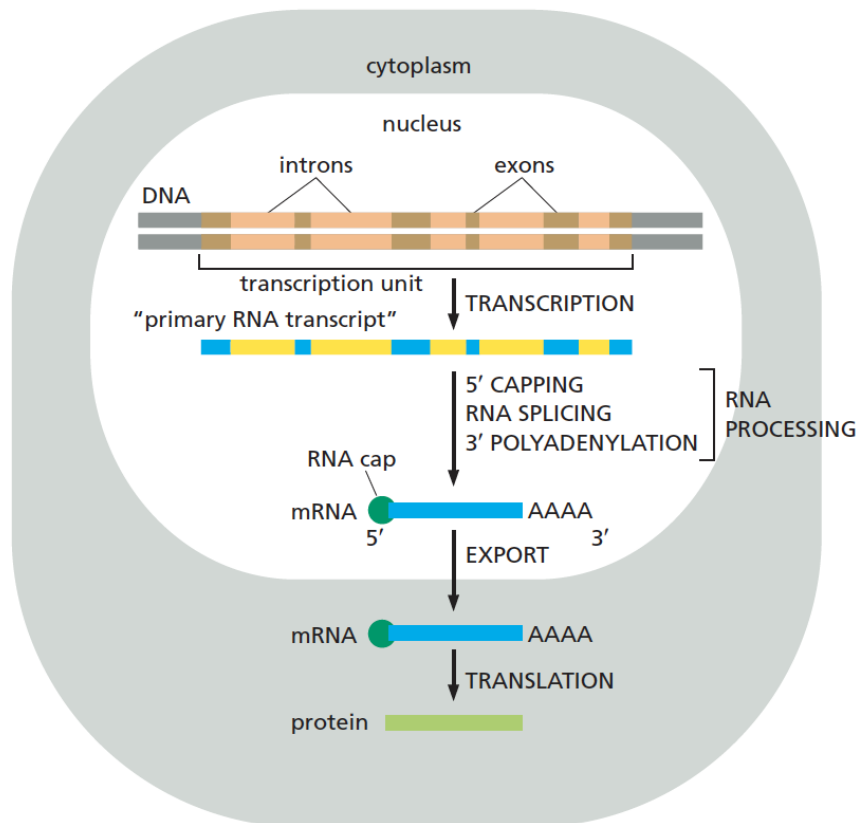


Figure 1.2: Steps in gene expression (Alberts et al., 2002).

1.1.2 Gene expression and regulation

Gene expression studies identify the genes that are expressed in cells, and then delineate genes that are differentially expressed in cells in different states or between different cell types. Molecular biology techniques, in particular transcriptome sequencing, has revolutionised biology in recent years and allows us to explore the regulation of physiological functions of a species at the gene level. Once the differentially expressed genes have been found, examining when a gene is expressed, and in which cell of the whole organism, can provide clues regarding gene function (Levine and Tjian, 2003). Gene regulation and changes in gene expression are essential to define the identity of cells during development and coordinate cellular activity throughout the cell's life.

Gene regulation is complex and occurs via several mechanisms, including transcriptional regulation (where and when genes are activated or silenced), post-transcriptional regulation (which exons remain after splicing), translation (where the rate of protein synthesis can be controlled), and RNA and protein stability (Lee and Young, 2013). Genome components

and the structure of chromatin play crucial roles in the regulation of gene expression at the transcriptional level, and include modified DNA bases and histones, remodelled chromatin structure and sequence variations which affect cell activity (Martin and Zhang, 2005; Kouzarides, 2007).

1.1.3 Epigenetics

In classical genetics, the nucleotide sequence contains the primary information, and changes in the nucleotide sequence influence phenotypes. As knowledge of genetics evolved, the concept of epigenetics was introduced by Conrad Waddington (1942) as a mechanism, in addition to the nucleotide sequence, that regulated the interaction between genes and phenotypes. Now, epigenetics refers to the study of potentially heritable changes that affect gene activity and expression but are not genetically encoded. Epigenetic mechanisms do not change the underlying DNA sequence of an organism (Weinhold, 2006).

Epigenetic changes are a critical facet of gene regulation that control development and are, in part, heritable. Epigenetic mechanisms include the recruitment of molecular processes, guided by small RNA molecules which could assist or degrade existing transcripts (Kim et al., 2009). Other epigenetic mechanisms affect chromatin structure and have been well studied, such as DNA methylation (adding methyl groups to the DNA molecule), histone modification (post-translational modifications of histones) and chromatin remodelling (dynamic modification of chromatin structure to affect the binding of proteins that regulate transcription) (Kim et al., 2009). Dosage compensation and genomic imprinting are two well-characterized processes that are epigenetically regulated and act during development by silencing regions of chromosomes or specific genes, as discussed below.

1.1.3.1 Dosage compensation

Dosage compensation in mammals is the process of balancing the expression of genes on sex chromosomes, between females (XX) and males (XY). Because of the evolutionary loss of genes from the Y chromosome and the presence of two X chromosomes in the female, there is a difference in the gene copy number between males and females, and hence

dosage compensation is needed to adjust the level of expression of the X-linked genes. X chromosome inactivation is the most important mechanism to balance the unequal genetic information carried between females and males. One of the X chromosomes in females is inactivated to ensure that the same number of X-linked genes are active in both sexes (Lyon, 1962). Both X chromosomes in female zygote are fully active (Epstein et al., 1978), then transcriptional silencing of one X chromosome occurs in early female embryonic development, and it remains inactive throughout subsequent cell divisions (Brockdorff, 2011). X chromosome inactivation (XCI) has two different forms: random and imprinted. After fertilization, paternally inherited X (Xp) is completely inactivated, this process called as imprinted XCI. Subsequently, Xp is reactivated in the inner cell mass of the blastocyst. Later, in the epiblast, either maternal or paternal X chromosome is transcriptionally silenced, this is called random XCI (Brockdorff and Turner, 2015). Imprinting differs from random XCI with respect to both the developmental timing and mechanism of action (Dementyeva et al., 2010). But both random and imprinted XCI are controlled by a locus on the X chromosome, known as the X Inactivation Centre, which includes the XIST gene (Borsani et al., 1991). XIST is a 17-kb long non-coding RNA that is expressed only from the inactive allele on female X (Brown et al., 1992; Clemson et al., 1996).

1.1.3.2 Genes that escape X inactivation

In placental mammals, the X chromosome contains ~900 genes, and most of the genes on the inactive X chromosome (Xi) are silenced (Tukiainen et al., 2017). But the Xi is not completely inactive (Carrel and Willard, 2005). New sequencing technologies have facilitated a detailed assessment of allelic expression and chromatin marks, to identify the genes that escape X inactivation (Tukiainen et al., 2017). Lyon's hypothesis (Lyon, 1962) is that X linked genes with Y gametologues always escape XCI. Although there are only 17 X-linked genes with Y gametologues in human, several studies have reported that ~20% of X-linked genes escape XCI in some tissues while ~12% of genes including X gametologues showed consistent inactivation in all somatic tissues (Carrel and Willard, 2005; Cotton et al., 2015; Tukiainen et al., 2017).

1.1.3.3 Genomic imprinting

Mammals are diploid organisms where both paternal and maternal genomes contribute to normal development. For most autosomal genes both parental alleles are expressed. But for some genes the allele inherited from only one parent is expressed, a phenomenon called imprinting. Most of the imprinted genes play key roles in fetal development. Failure to establish correct imprinting perturbs neonatal growth which could result in e.g., neurological disorders such as Prader-Willi syndrome (Barlow and Bartolomei, 2014).

The first examples of genomic imprinting discovered in mammals were three imprinted genes, *IGF2R*, *IGF2*, *H19*, identified in mouse. *H19* was found to be a maternally expressed gene (Bartolomei et al., 1991) and *IGF2* a paternally expressed gene (Ferguson-Smith et al., 1991). *IGF2R*, a fetal growth factor receptor, was then shown to be imprinted in many mammalian species (Zemel et al., 1992; Dindot et al., 2004). To date, about 151 imprinted genes have been identified in mouse and ~100 in human (<http://igc.otago.ac.nz>; http://www.har.mrc.ac.uk/research/genomic_imprinting). Most studies of imprinting have been carried out in human and mouse, so the investigation of imprinted genes in other mammals such as cow, pig and sheep have tended to focus on characterizing the genes previously known as imprinted in human and/or mouse (Thurston et al., 2008; Bischoff et al., 2009; Chen et al., 2016). As a result, the number of imprinted genes identified in large mammals, including sheep, cattle and pig, is much lower than in human and mouse.

1.2 Mammalian development

1.2.1 Mammalian prenatal development

A wide range of phenotypes are largely determined prenatally. Stimuli or insults in the critical period of prenatal development can have lifetime consequences.

Prenatal development can be divided into three stages: the germinal stage, the embryonic stage, and the fetal stage (Berk, 2000). The germinal stage starts with the fertilisation of a mature oocyte by a spermatozoon, followed by the formation of the zygote, which is the

first diploid cell formed following fertilisation (Oestrup et al., 2009). The zygote divides into over 100 cells to become the blastocyst, which is made up of three germ layers; the endoderm; mesoderm and ectoderm. Cells in these three layers will give rise to different tissues in the organism (MacCord, 2013).

The blastocyst migrates to the uterus and attaches to the uterine wall, a process known as implantation. Unlike other vertebrates (such as amphibians), mammalian embryos must be implanted into the uterine wall for development. In human, the period between the 3rd and 8th week following fertilisation is called the embryonic period as cells continue dividing, and the germ layers develop into tissues and organs. The ectoderm eventually forms epithelial tissues such as skin, hair and also the nervous system (Grubb, 2006). The mesoderm develops into muscle, the skeletal system and connective tissues. The cells in endoderm give rise to certain organ such as liver, lung and stomach. Fetal limb development begins at day 28 of pregnancy (Barham and Clarke, 2008) and the other organs form during this early stage of pregnancy include the pancreas, liver, adrenal gland, lung, thyroid, spleen, brain, thymus and kidney (Schmidt-Rhaesa, 2007).

After the embryonic period, about week 9 following fertilization in humans, cell differentiation and organ formation are mostly complete and the embryo enters the next developmental stage and becomes a fetus (Carlson, 2018). This marks the start of a rapid growth phase, as well as the ongoing differentiation of organ systems established in the embryonic period. The brain continues to grow and develop, the urogenital system differentiates between males and females, and the endocrine and gastrointestinal tract begins to function. Muscle and adipose tissue formation occur at mid-gestation (Du et al., 2010). Organ and body structures continue to develop from this stage until birth.

Studying the causes of congenital anomalies in fetal development is important in developmental biology. Analysing gene pathways that are related to developmental anomalies is one way to help us develop new approaches for the prevention and treatment of diseases. However, considering the difficulty of sample collection and ethical issues working with human fetal material, mouse models have been widely used in embryonic and fetal studies (Mayer and Joseph, 2013). Although mouse embryos are easy to access, the development

of human and mouse is different. The average mouse gestation is only 21 days and the developmental stage of the newborn mouse is equivalent to a human week 10 fetus (Otis and Brent, 1954). The average length of human gestation is 280 days, similar to cattle where it ranges from 279 to 287 days (Livesay and Bee, 1945). This makes cattle a better choice to study mammalian prenatal development of relevance to humans.

1.2.2 Sexual dimorphism during development

Males and females of many mammalian species differ in their physical appearance, including body size and weight. Human male infants have a higher birth weight than female infants, while human adults show sex-specific health and behavioural differences (Regitz-Zagrosek, 2012). For example, compared to female, male is more likely to have cardiovascular disease, schizophrenia, and Parkinson's disease (Nojiri et al., 2019). In the livestock industry, sex differences affect breed and sex selection. For example, dairy farmers prefer female calves because they can produce milk, whereas male calves are preferred by beef producers as male always have large carcass size.

At the molecular level, sexually dimorphic phenotypes are the result of complex genetic architecture, which control the production of gonadal sex hormones that drive prenatal development of many of the phenotypic differences. Sexual dimorphism typically appears after gonadal differentiation (Fujimoto et al., 2010). Sex-related hormones secreted from differentiated gonads can affect the expression of sex-specific phenotypes. In addition, these hormones affect the sex-specific behaviours by changing brain function in a process of masculinization and feminization (Ngun et al., 2011). The differences of fetal growth rates between male and female have long been recognized (Lubchenco et al., 1963). Some sex differences of embryos are seen early before the gonadal development, this suggests the involvement of factors other than gonadal hormones (Cook and Monaghan, 2004). Studying sexual dimorphism in non-gonadal somatic tissue during prenatal development will help us to identify the potential candidates that drive sex specific differences.

1.3 Next Generation Sequencing technology

1.3.1 Short and long read sequencing

Determining the genomic sequences has contributed immensely to our views on genome structure and function. Over 50 years ago, methods of primer extension were used to determine a sequence of the bacteriophage lambda (Wu and Kaiser, 1968). Maxam and Gilbert used chemical cleavage successfully to determine the sequence of the lactose-repressor binding site in following years (Gilbert and Maxam, 1973). In the mid-1970s, Frederick Sanger and colleagues developed a technique that used chain termination dideoxy nucleotides for DNA sequencing (Sanger et al., 1977). Although Sanger sequencing has limitations, especially low sample throughput, it was the method that transformed biology by facilitating the accurate sequencing of DNA. In 1987, an automated, Sanger-based DNA sequencer was introduced by Applied Biosystems, enabling the sequencing of several DNA fragments in parallel using dye termination chemistry and electrophoresis (Watson and Cook-Deegan, 1991). This automated DNA sequencer was the workstation at the heart of the sequencing of the first draft of the human genome (Collins et al., 2003). Sanger sequencing remains a most accurate technology and is still in use today. Subsequently, so-called “Next Generation Sequencing” (NGS) platforms were developed for high throughput sequencing of DNA and RNA. These were “game changers” that revolutionized genetics. In the last 10 years, NGS platforms have led to an exponential increase in our knowledge of genetic variation and have facilitated fast and accurate diagnostics in human clinical studies (Ng et al., 2010).

The Illumina NGS platforms have been the most widely used for short-read sequencing of DNA and RNA—the latter after conversion to cDNA (Meyer and Kircher, 2010). DNA is randomly fragmented to ~300-400 base pair fragments, followed by 5' and 3' adapter ligation to attach sequencing adaptors. Each DNA molecule is amplified to form clusters with the same sequence. The clusters are then sequenced base by base at high accuracy. A major improvement in NGS technology is the paired-end (PE) sequencing which involves sequencing the DNA fragments in a library from both 5' and 3' ends and aligning reads in pairs. Sequences aligned as read pairs have a higher mapping accuracy (Illumina, 2021).

Current instruments produce TB of data with paired-end (PE) reads of up to 150bp each (Liu et al., 2012). Short-read sequencing is now widely used as costs have continued to fall.

Sanger and NGS technologies deliver DNA reads up to 1000 bases. These short reads do not allow the analysis of complex genomic loci, genome rearrangements, nor longer repetitive elements. This results in incomplete genome assemblies, especially for large and complex genomes. Variant phasing over large distances (haplotyping) is also very poor. PCR amplification for sequencing will produce artefacts and exclude the detection of natural base modifications, such as methylation (Acinas et al., 2005). Several of these disadvantages have been overcome by third-generation sequencing technologies: long-read sequencing (Amarasinghe et al., 2020).

Long-read sequencing (LRS) allows for the generation of much longer (>10,000bp) sequence reads with lengths of up to 100,000bp not uncommon (Lee et al., 2016). LRS technology directly sequences single molecules of DNA in real-time, which produces much longer reads than those from SRS, but this does require careful DNA preparation to avoid fragmentation. Sequencing long reads from native DNA with little or no enzymatic treatment has enabled the resolution of tandem repeat expansions and GC-rich regions with improved accuracy (Mantere et al., 2019). To date, there are two leading LRS producers: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore), the major issue with LRS is that it remains much more expensive than SRS.

1.3.2 Transcriptome sequencing

Transcriptome sequencing has been applied widely including the identification of the complete spectrum of the entire transcriptome and RNA quantification. Most transcriptome studies compare gene expression levels across various treatments, tissues, types of individual (breeds or genotypes) or physiological states (Moridi et al., 2019; Pareek et al., 2019). Gene expression studies have used low-throughput sequencing methods such as quantitative polymerase chain reaction (qPCR), which tend to be limited to measuring a few transcripts. In 1995, profiling of a whole transcriptome was carried out

using hybridization-based Microarray technologies, the first relatively low cost, “high-throughput” approach (Schena et al., 1995). However, only genes with designed probed can be studied in microarray datasets. Subsequently, the direct sequencing of transcripts by SRS technologies (RNA-seq) has become the preferred method, as it can identify transcripts that are not present on microarrays.

Ribosomal RNA (rRNA) is one of major types of non-coding RNA that is isolated from cells and plant and animal tissues, comprising >80-90% of total cellular RNAs (O’Neil et al., 2013). The highly abundant Ribosomal RNA needs to be removed from total RNA to allow efficient detection of coding and non-coding transcripts. Standard approaches include polyadenylated RNA (PolyA+) selection, the use of oligo (dT) primers to select mature polyA+ transcripts, and rRNA depletion i.e., a hybridization-based method that remove rRNAs which is bound to paramagnetic beads (Zhao et al., 2018).

The read depth offered by next-generation sequencing enables transcripts with relatively low level of expression to be quantified. In theory, RNA-Seq can be used to study levels of different isoforms generated by alternative splicing, even when the isoforms are not fully annotated. In practice, although numerous tools have been developed to detect splicing junctions for short read sequences such as DiffSplice (Hu et al., 2013) and JunctionSeq (Hartley and Mullikin, 2016), only two methods rSeqDiff (Shi and Jiang, 2013) and rMATs (Shen et al., 2014) have been able to detect differentially spliced isoforms. However, because of the limitation of the short reads, the accuracy of these methods is still questionable (Ding et al., 2017).

The median length of human gene transcripts is about 2.5k bp, therefore the 2×150bp read length is insufficient for good transcriptome reconstruction. For most of the genes, reads from only one isoform are typically identified (Steijger et al., 2013). A study of RNA-seq quantification errors also suggested that RNA-Seq is unable to measure expression accurately within gene families, which are often enriched in human disease (Robert and Watson, 2015). For the identification of isoforms of the same gene or analysis of gene families, the use of LRS is preferable.

In 2015, Pacific Biosciences developed long-read sequencing for RNA (Gonzalez-Garay,

2016). This method, known as Iso-Seq, is becoming widely used for transcript identification. Iso-Seq is capable of identifying novel transcripts and isoforms with extraordinary precision, because reads can be 10kb or longer and the quality of full-length reads can reach >99% accuracy. In a study of mRNA transcripts in red clover, Iso-Seq was used to identify splice isoforms. As a result, a total of 29,730 isoforms from known genes and 2,194 isoforms from novel genes were identified, adding important information to red clover transcriptome expression profiling (Chao et al., 2018).

1.4 Bioinformatics methods

With ongoing developments in sequencing technologies, the need for tools to analyze the increasingly large and complex data sets became apparent. These fall under the aegis of Bioinformatics, which combines biology, computer science and statistics for biological studies. Analysis of genes, transcripts, proteins and epigenetic features play a central role. Most DNA sequencing techniques produce sequences that can be used to assemble complete gene or genomes. Assembling the sequences is very complicated for larger genomes such as mammalian genomes: it is computationally difficult, time-consuming and incomplete. It may take many iterations and a range of data types to assemble and orient sequences correctly. Notwithstanding, a complete and high-quality reference genome is key for downstream analysis.

Transcriptomes are used to explore the functional elements of the genome, reveal the molecular differences between cells and tissues, and to understand biological processes, such as development and disease. A large range of methods and software tools have been developed to analyze transcriptome data. The steps of genome assembly and bioinformatics processes in transcriptome analysis are discussed below.

1.4.1 Steps of genome assembly

1.4.1.1 *De novo* sequence assembly using long read data

The goal of genome assembly is to assemble a complete genome using highly accurate sequencing reads. However, reads from repeated regions of the genome lead to ambiguous

assembly. This problem has been partially resolved using single-molecule sequencing, which can produce much longer reads (Gordon and Hannon, 2010). However, single-molecule sequencing is less accurate than short read technologies (Eid et al., 2009). Therefore new alignment methods are required that are able to detect and correct sequencing errors than those used for short read sequences. Most genome assemblies are now generated using long reads generated by PacBio and Oxford Nanopore technologies. Assemblers written to assemble sequences from long-read data include Falcon-unzip (Chin et al., 2016) and Canu (Koren et al., 2017). Falcon-unzip is a PacBio in-house designed diploid assembler that is only suitable for the PacBio long-reads data. The error-corrected reads are assembled into contigs by FALCON and then contigs with heterozygous SNPs are found that identify the haplotype of each read. Next, the phased reads are used to generate primary contigs and fully phased haplotigs which represent divergent haplotypes (Chin et al., 2016). Canu is a common and efficient assembler using an adaptive k-mer weighting strategy and an automated error rate estimation feature for both PacBio and Nanopore data. It is also the first method that is able to assemble complete haplotypes from a heterozygous diploid genome and accurately reconstruct structurally heterozygous alleles (Koren et al., 2017; Koren et al., 2018).

1.4.1.2 Scaffolding

Although the contigs generated from long reads tend to be high quality, their length is still too short to assemble into chromosomes. Scaffolding is a necessary step in genome assembly which links a contig into a scaffold. Genome scaffolding tools use Hi-C reads and/or genetic/physical maps as guide to link scaffolds. But the challenge of scaffolding is how to deal with the high repetitive content of genomes and how to fix mis-assembled contigs. Hi-C (Lieberman-Aiden et al., 2009) and optical maps (Mak et al., 2016) have been extensively used in many large-scale mammalian genome assembly projects such as goat (Bickhart et al., 2017), water Buffalo (Low et al., 2019) and cattle (Low et al., 2020). Hi-C was invented to study the spatial organization of DNA in a cell and has been also used for scaffolding an assembly (Low et al., 2020). Alignment of HiC generated paired end sequences to a reference genome enables physical contacts in chromatin to be identified

(Pal et al., 2019). However, since Hi-C cannot provide an accurate orientation of contigs, it can be used to order long contigs while the size of gaps between contigs cannot be estimated. It works best in linking longer contigs. Salsa (Ghurye et al., 2017) and 3d-DNA (Dudchenko et al., 2017) use Hi-C data sets for scaffolding.

Optical mapping is a new sequencing-free technology that fluorescently labels DNA fragments at specific sequence motifs then measures the length between these labelled motifs by passing the fragments through a nanochannel array which is imaged to generate single molecule DNA maps. These maps then can be used order contigs and also estimate the size of gaps between contigs in a whole-genome assembly. It works well in linking both long and short contigs. A combination of Hi-C and optical mapping is now widely used to assemble chromosome scale scaffolds such as for goat (Bickhart et al., 2017) and cattle (Low et al., 2020).

1.4.2 Bioinformatic processes in transcriptome analysis

1.4.2.1 Quality control and alignment

The first step in transcriptome analysis is quality assessment of the reads. Reads are processed to increase the likelihood of the success of subsequent analyses. Software such as FastQC (Andrews et al., 2010) and ChiLin (Qin et al., 2016) are frequently used to assess the overall quality of the reads, the distribution of base-call quality scores, presence of no-calls, read duplications, over-represented sequences etc. When the sequenced fragments are shorter than the read length, adjoining adaptors will also be present, and must be removed. Programs such as Cutadapt (Martin, 2011) and fastx (Gordon and Hannon, 2010) are designed to do this. Once the reads have been quality checked, trimmed and poor-quality reads removed, they are ready to be mapped to a reference genome to determine the location from which the reads originated. Alignment of reads to the reference genome is a huge topic and a detailed analysis of alignment methods is beyond the scope of this introduction. Alignment must allow for mismatches between the sample and reference sequences. Most cDNAs map discontinuously to the reference genome, because introns have been spliced out, and alignment must take this into account. Several programs have

been developed for this purpose such as *hisat2* (Kim et al., 2015), *star* (Dobin et al., 2013) and *subread* (Liao et al., 2013). In species where a reference genome is not available, programs such as *salmon* (Patro et al., 2017) and *Kallisto* (Bray et al., 2016) carry out the *de novo* assembly of reads to identify transcripts.

1.4.2.2 Variant Calling

Variant detection software typically aligns reads from a sample to a reference genome and determines where bases differ. A threshold is then set such that if a difference occurs sufficiently frequently a variant is called. Two frequently used variant calling programs are *GATK* (McKenna et al., 2010) and *freebayes* (Garrison and Marth, 2012). These programs can detect single nucleotide variants (SNVs) and small indels (1 - 10 bp) using either single-sample or multiple-sample variant calling. *GATK* carries out local *de novo* assembly of haplotypes and then uses a Bayesian model for genotyping and variant calling. However, calling short indels is less accurate than calling SNPs, and *GATK* cannot call long indels which are more than 100 bp. *Freebayes* has a higher sensitivity in detecting variants with less than 5% minor allele frequency than other methods (Sandmann et al., 2017). It also requires less compute time than *GATK* (Hwang et al., 2015). In general, calling variants from RNA is not as accurate as from DNA, as cDNA contains errors produced by the relatively low-fidelity Reverse Transcriptase, and changes due to RNA-editing. The common variant calling tools do not attempt to address these.

1.4.2.3 Differential gene expression

RNA-seq studies compare gene expression levels within and between samples, from various treatments or between samples of different types. Several bioinformatic methods have been designed to detect differentially expressed (DE) genes or transcripts in transcriptome data. Each RNA-seq study requires at least 3 biological replicates (and preferably 5 or more) in each group to generate sufficient statistical power (Schurch et al., 2016). Counts of reads are generated per gene or transcript. They are not normally distributed but tend to have a negative binomial (NB) distribution, and this has been used to estimate variability among biological replicates. R packages using NB models include *edgeR* (Robinson et al.,

2010), baySeq (Hardcastle and Kelly, 2010) and DESeq2 (Love et al., 2014). However, NB dispersions have the limitation of not allowing for gene-specific variation and cannot effectively handle genetic variation among biological replicates within a group (Law et al., 2014). Limma-voom (Law et al., 2014), an R package for RNA-Seq DE analysis, applies normal distribution-based statistical methods to read counts by estimating the mean-variance relationship using log-transformed counts from biological samples. This method is fast and works well even when samples are of low-quality (Liu et al., 2015).

1.4.2.4 Pathway analysis

Once DE genes have been identified, attention turns to establishing a biological context. This can be done by exploring networks involving DE genes, in order to understand their role in specific biological processes or molecular functions. A common approach is to compare a list of DE genes to annotated databases of genes. Gene Ontology and KEGG (<http://geneontology.org/>; <https://www.genome.jp/kegg/ko.html>) are two public resources commonly used to find pathways enriched for DE genes. After comparing the genes to the database, several software packages visualize genes on pathway maps or rank functional categories according to the co-occurrence of genes in a gene list. These include WebGestalt (Wang et al., 2013), DAVID (Huang da et al., 2009), camera (Wu and Smyth, 2012) and GSEA (Subramanian et al., 2005).

In summary, with the rapid advances in sequencing technologies, bioinformatics methods to analyse biological data sets in many species have been developed. However, most of existing software are only able to analyse one particular type of sequencing data. As the volume and complexity of data increases, it is becoming increasingly necessary to customize existing software and create new algorithms/pipelines to combine data from different sequencing platforms to investigate and aid interpretation.

1.5 Introduction to cattle

1.5.1 Cattle domestication

The relationship between humans and cattle has existed for more than ten thousand years (Zeder, 2015). Cattle play an important role as a producer of human daily needs such as food and clothes. In return, humans provide the essential requirements for cattle, including feed, water and care. The economic importance of cattle has resulted in the development of specialized dairy and beef breeds (Schibler and Schlumbaum, 2007). Researchers have proposed various theories on the domestication of cattle, and the time that it occurred, based on archaeological evidence (Helmer et al. 2005). With the improvements in molecular biology techniques, DNA sequence-based approaches can now produce genetic data to corroborate or refute archaeological inferences, and also add information on geographic origin and relationships among populations. Identifying the history and origins of domestic cattle may also help to reveal potential sources of genetic diversity, which can be used to improve livestock adaptation and agricultural production.

Genetic studies have shown that the first cattle were domesticated from wild aurochs (known as *Bos primigenius*) around 10,000 B.P. Since domestication cattle have spread across the world. Modern cattle breeds fall into two subspecies: taurine cattle (*Bos taurus taurus*), the ancestors of which were domesticated in the Fertile Crescent from *Bos primigenius* during the Neolithic period (Ajmone-Marsan et al., 2010; MacHugh et al., 2017; Pitt et al., 2019), and indicine cattle (*Bos taurus indicus*), whose ancestors were domesticated in the Indus Valley from *Bos primigenius* approximately 1,500 years later (Loftus et al., 1994).

1.5.2 *Bos taurus taurus* and *Bos taurus indicus*

After Fertile Crescent domestication, taurine cattle dispersed quickly into Southern Europe and subsequently into all parts of Europe (Pitt et al., 2019). Taurine cattle generally have short hair, a flat back and are more adapted to temperate regions. Indicine cattle spread across the Indian sub-continent and then to the East coast of Africa and to South-East Asia (Ajmone-Marsan et al., 2010). Indicine cattle have a pronounced back hump, long

drooping ears, and dewlap folds that provide a greater skin surface area for losing heat (Hansen, 2004), making them better adapted to tropical climates.

Taurine cattle are often managed by breed societies that oversee intense selection for production traits, contrasting with indicine cattle that have been less intensively selected for productivity. As indicine cattle have generally not been intensely selected, and thus not had their genetic diversity substantially reduced, they may show more ability to adapt to adverse conditions, such as local disease and hot climates (Zeng et al., 2019).

Crossbreeding programs in tropical environments frequently use indicine and taurine breeds, to take advantage the genetic differences between them and the resulting hybrid vigor. For example, Brahman-Hereford heifers have higher post-weaning live weight and daily weight gain than pure Hereford heifers in both temperate and subtropical environments (Arthur et al., 1999). Many composite taurine-indicine cattle breeds have been developed over the last century (Frisch and Vercoe, 1977) and are now extensively used to increase the productivity of cattle in Australia, especially in Northern Australia.

1.5.3 Genetic variation

Studying the genetic differences among populations to identify variants that control phenotypic differences will advance knowledge and accelerate selection for desired traits. Single-Nucleotide Polymorphism (SNP) markers discovered by genome sequencing have been extensively used to analyse genetic variation and identify its association with phenotypic diversity in a wide range of species (Hiremath et al., 2012). Many bovine SNP markers have been identified and used to create genotyping panels for the genetic analysis of cattle populations (Van Tassell et al., 2008). SNP panels at different SNP densities have been created for use in cattle e.g. Illumina GGP-LD 30K chip and the Versa50K chip. The 770k Illumina BovineHD chip is the most comprehensive chip available for cattle genome wide genotype study, with 777,962 SNPs that have been validated on more than 28 *Bos taurus taurus* and *Bos taurus indicus* breeds.

SNP markers have been used to explore the antecedents and relationships among modern cattle populations. A recent study analyzed more than 3,000 cattle samples belonging

to 180 taurine and indicine breeds with ~54,000 SNPs, to elucidate population structure by approximate Bayesian computation (Pitt et al., 2019). There was a clear separation between Eurasian taurine, African taurine and the indicine breeds (Figure 1.3), supporting the hypothesis that modern cattle resulted from two separate domestication events (Pitt et al., 2019).

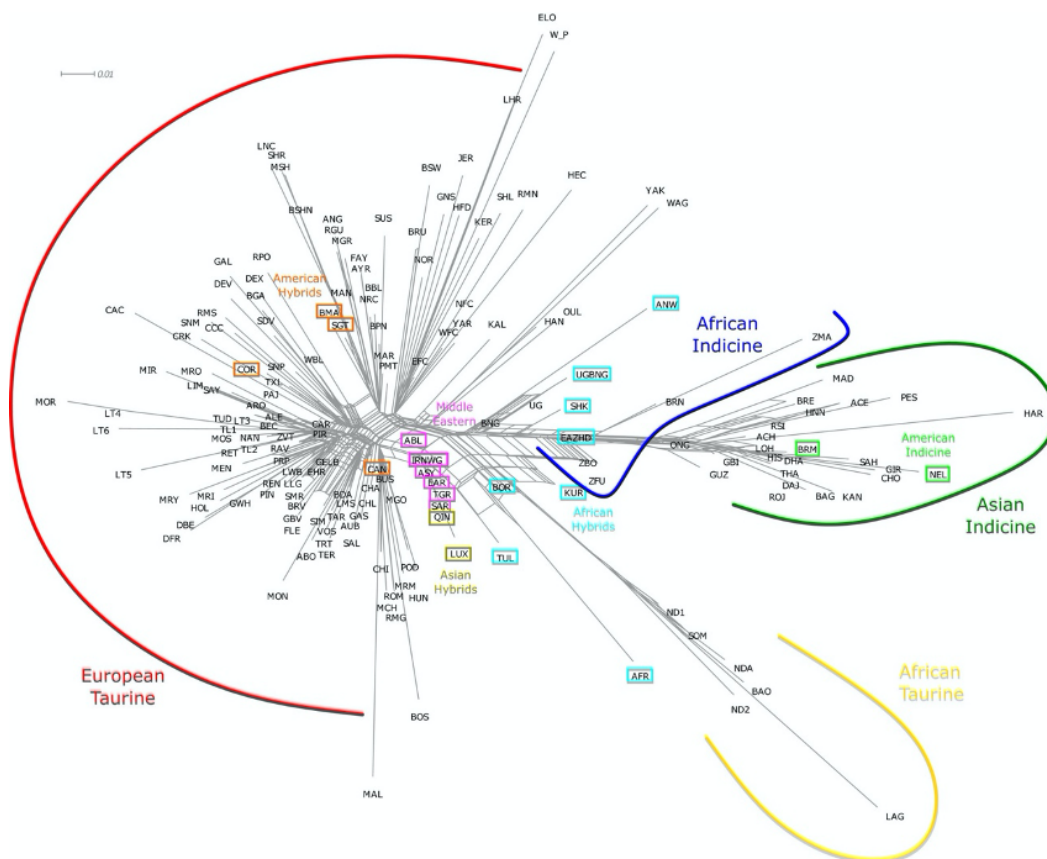


Figure 1.3: Neighbour-net analysis for 174 populations of taurine and indicine cattle (Pitt et al., 2019).

Linkage Disequilibrium (LD) maps, also known as allelic association maps, can be created using these SNP panels (McKay et al., 2007). They can be used in quantitative trait locus (QTL) mapping, a method that uses historical LD to link phenotype to genotypes and explore the genetic basis of economically important traits. It also enables us to explore the genetic diversity of cattle populations and to identify genomic regions with selective sweeps (McKay et al., 2007). However, most of the SNP markers used to generate LD maps were derived by taking sequence of *Bos taurus taurus* animals and aligning it to a *Bos taurus taurus* reference genome, and so may not be polymorphic in *Bos taurus indicus* cattle.

Structural Variation (SV), which refers to deletions, duplications, inversions and translocations greater than about 1 kilobase (kb), accounts for much more genetic variation than SNP diversity (Iafate et al., 2004). Copy Number Variations (CNV) are one of the important subsets of structural variation, comprising of deletions and duplications. SVs that span gene coding regions have been shown to affect a range of traits such as milk production, fertility, and disease resistance (Liu et al., 2010; Medugorac et al., 2012; Kadri et al., 2014). For example, in Angus cattle, 297 CNVs were found to be associated with parasite resistance which is overlapped with 437 genes that were enriched in immune function (Hou et al., 2012). However, it is difficult to assay CNVs with currently available routine genotype platforms, so they tend not to be used in association studies. Such studies would require many samples as small sample sizes will not be sufficiently powered to find subtle differences in associations with traits. Only a small number of cattle CNVs have been characterised. Therefore, it is difficult to understand the phenotypic impact of the individual CNVs.

1.5.4 Cattle genome assembly

Accurate genome assemblies are an invaluable resource for research. Gene expression analyses rely on the availability of a high-quality reference genome to map reads, create accurate gene models and discover features such as alternative splicing. Therefore, much effort has been devoted to creating ever-more complete and accurate reference genomes for many species. There are currently 60 annotated mammalian reference genomes available in the NCBI assembly database. This is a comprehensive database providing chromosome-level DNA sequences and also contains gene information for each genome assembly (<https://www.ncbi.nlm.nih.gov/assembly/>).

The first draft genomic sequence of the cow (Btau 4.0) was generated by the International Bovine Genome Consortium (Bovine Genome et al., 2009). This assembly used a combination of bacterial artificial chromosome (BAC) end sequencing and whole-genome shotgun sequencing (WGS) to assemble reads primarily from a Hereford cow called Dominette (*Bos taurus taurus*), but also included sequences from other sources, including BAC clones from the Dominette's sire Domino. Radiation hybrid (RH), genetic linkage

and cytogenetic maps were used to order the sequences and assign the short scaffolds to chromosomes. As most of the sequence was from a cow there was no Y chromosome assembled.

The Center for Bioinformatics and Computational Biology at the University of Maryland used the same raw sequence data to generate an alternative assembly, UMD2, which was also published in 2009 (Zimin et al., 2009). This assembly used the same reads as Btau 4.0, combined with independent marker data from the International Bovine BAC Consortium (IBBMC) clone fingerprint maps and a composite linkage/RH map (Snelling et al., 2007). One of the most important differences between the Btau 4.0 and UMD 2.0 assemblies is the size of the X chromosome, whereas the length of X chromosome in the UMD2.0 assembly is ~136 Mbp, the Btau 4.0 assembly it is only 83 Mbp.

The first indicine (*Bos taurus indicus*) genome sequence, bos_indicus 1.0, was generated using the SOLiD sequencing platform (Canavez et al., 2012). The reads were aligned to the 2 *Bos taurus taurus* genome assemblies (Btau 4.0 and UMD2.0) creating a pseudo-assembly. This assembly has not been used widely as the alignment-based assembly will have lost some indicine features. A full high-quality *de novo* Indicine genome sequence assembly has been required for some time, to study the genetic differences between taurine and indicine subspecies.

With the development of third generation, long-read sequencing technologies (Dijk et al., 2018), alongside chromosome conformation capture techniques, such as Hi-C (Belton et al., 2012), it has become possible to generate very high-quality genome assemblies relatively quickly and cheaply. In 2017, a *de novo* goat genome was the first diploid vertebrate genome assembly generated using PacBio long-read sequencing combined with Hi-C and optical mapping data (Bickhart et al., 2017). This was the most contiguous genome available at that time. The most recent cattle genome assembly, ARS-UCD1.2, became available in 2018 (RefSeq assembly, GCF_002263795.1). This assembly was generated using the same third-generation sequencing technologies as goat genome and was, again, based on reads from Dominette, the Hereford cow used in the first bovine genome assemblies. Because Dominette is female, the assembly still does not include a

cattle Y chromosome.

Sequencing the mammalian sex chromosomes, and the Y chromosome in particular, is especially difficult because of abundant repetitive sequences (Kuderna et al., 2019). Human, chimpanzee, mouse, pig and horse are the only mammalian species with well-characterized Y sequences (Skaletsky et al., 2003; Hughes et al., 2010; Soh et al., 2014; Skinner et al., 2016; Janecka et al., 2018). The current cattle X chromosome contains regions with duplicated sequences but is much more complete than the Y chromosome assembly (Liu et al., 2019). The only available cattle Y chromosome is from btau_5.0.1, which used five BAC libraries from at least three different breeds as the resource for the assembly (Rozen et al., 2006). The composite nature of the starting materials has likely contributed to errors in the assembly of duplicated and repeated sequences.

1.6 Research aim

My project is focused on the integration of computation and biology to address biological questions in bovine fetal development. Fetal development is a process of forming organs and tissues and is mediated by gene regulation. Determining the sex-specific and genotype-specific patterns of gene expression will provide information about normal development. Identification of alterations in these patterns of gene expression that are associated with specific phenotypes may also help to elucidate the underlying molecular mechanisms. Uncovering qualitative and quantitative changes in gene expression and understanding the mechanisms that regulate gene expression in bovine fetal development will contribute new knowledge. To carry out this work a well-assembled cattle reference genome is essential, but the cattle reference genome available when I began my study was incomplete and in particular missing a Y chromosome. Genes located on Y chromosome have been reported to play very important functions in germ cell development and are related to male fertility (Lahn and Page, 1997). A well assembled Y chromosome will enable us to study roles of Y genes in sex difference in phenotypes.

For my studies I used material from purebred and reciprocal cross individuals of the two cattle sub-species, *Bos taurus indicus* and *Bos taurus taurus*. Purebred and hybrid

fetuses have considerably different phenotypes e.g., for weights and growth rates, with pronounced sex effects in both sub-species (Xiang et al., 2013). A previous study on bovine fetal bone using the same experimental animals suggested that there are differential effects of paternal and maternal genomes combined with fetal sex effects (Xiang et al., 2014).

Specifically, my study addressed parental genome effects and sex effects in five tissues (brain, liver, lung, muscle, placenta) in pure and cross-bred bovine fetuses by:

- (i) Assembling, annotating and comparing bovine sex chromosomes for *Bos taurus indicus* and *Bos taurus taurus* cattle, which is addressed in thesis § 2.
- (ii) Identifying candidate genes that drive sexual dimorphism and investigating the relationship between expression levels of these genes and fetal phenotypic differences in mid-gestation, which is addressed in thesis § 3.
- (iii) Identifying differentially expressed genes within and between fetal tissues of pure-bred and cross-bred cattle, which is addressed in thesis § 4.

References

- Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F., 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16s rRNA clone libraries constructed from the same sample. *Applied and environmental microbiology*, 71(12), pp.8966–8969.
- Ajmone-Marsan, P., Garcia, J.F., and Lenstra, J.A., 2010. On the origin of cattle: how aurochs became cattle and colonized the world. *Evolutionary anthropology: issues, news, and reviews*, 19(4), pp.148–157.
- Alberts, B., Johnson, A., Lewis, J., Walter, P., Raff, M., and Roberts, K., 2002. *Molecular biology of the cell 4th edition: international student edition*. Routledge.
- Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1), pp.1–16.
- Andrews, S. et al., 2010. *Fastqc: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- Arthur, P., Hearnshaw, H., and Stephenson, P., 1999. Direct and maternal additive and heterosis effects from crossing bos indicus and bos taurus cattle: cow and calf performance in two environments. *Livestock production science*, 57(3), pp.231–241.
- Barham, G. and Clarke, N.M., 2008. Genetic regulation of embryological limb development with relation to congenital limb deformity in humans. *J child orthop* [Online], 2(1), pp.1–9.
- Barlow, D.P. and Bartolomei, M.S., 2014. Genomic imprinting in mammals. *Cold spring harb perspect biol* [Online], 6(2).
- Bartolomei, M.S., Zemel, S., and Tilghman, S.M., 1991. Parental imprinting of the mouse h19 gene. *Nature* [Online], 351(6322), pp.153–5.

- Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J., 2012. Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods* [Online], 58(3), pp.268–76.
- Berk, L.E., 2000. *Child development*. Boston: Allyn and Bacon.
- Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., Sullivan, S.T., Burton, J.N., Huson, H.J., Nystrom, J.C., Kelley, C.M., Hutchison, J.L., Zhou, Y., Sun, J., Crisa, A., Ponce de Leon, F.A., Schwartz, J.C., Hammond, J.A., Waldbieser, G.C., Schroeder, S.G., Liu, G.E., Dunham, M.J., Shendure, J., Sonstegard, T.S., Phillippy, A.M., Van Tassel, C.P., and Smith, T.P., 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat genet* [Online], 49(4), pp.643–650.
- Bischoff, S.R., Tsai, S., Hardison, N., Motsinger-Reif, A.A., Freking, B.A., Nonneman, D., Rohrer, G., and Piedrahita, J.A., 2009. Characterization of conserved and nonconserved imprinted genes in swine. *Biol reprod* [Online], 81(5), pp.906–20.
- Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H.F., Avner, P., and Ballabio, A., 1991. Characterization of a murine gene expressed from the inactive x chromosome. *Nature* [Online], 351(6324), pp.325–9.
- Bovine Genome, S., Analysis, C., Elisk, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., Guigo, R., Hamernik, D.L., Kappes, S.M., Lewin, H.A., Lynn, D.J., Nicholas, F.W., Reymond, A., Rijnkels, M., Skow, L.C., Zdobnov, E.M., Schook, L., Womack, J., Alioto, T., Antonarakis, S.E., Astashyn, A., Chapple, C.E., Chen, H.C., Chrast, J., Camara, F., Ermolaeva, O., Henrichsen, C.N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S.M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J.M., Gerlach, D., Elhaik, E., Graur, D., Reese, J.T., Edgar, R.C., McEwan, J.C., Payne, G.M., Raison, J.M., Junier, T., Kriventseva, E.V., Eyraas, E., Plass, M., Donthu, R., Larkin, D.M., Reecy, J.,

- Yang, M.Q., Chen, L., Cheng, Z., Chitko-McKown, C.G., Liu, G.E., Matukumalli, L.K., Song, J., Zhu, B., Bradley, D.G., Brinkman, F.S., Lau, L.P., Whiteside, M.D., Walker, A., Wheeler, T.T., Casey, T., German, J.B., Lemay, D.G., Maqbool, N.J., Molenaar, A.J., Seo, S., Stothard, P., Baldwin, C.L., Baxter, R., Brinkmeyer-Langford, C.L., Brown, W.C., Childers, C.P., Connelley, T., Ellis, S.A., Fritz, K., Glass, E.J., Herzig, C.T., Iivanainen, A., Lahmers, K.K., Bennett, A.K., Dickens, C.M., Gilbert, J.G., Hagen, D.E., Salih, H., et al., 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* [Online], 324(5926), pp.522–8.
- Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L., 2016. Near-optimal probabilistic rna-seq quantification. *Nat biotechnol* [Online], 34(5), pp.525–7.
- Brockdorff, N. and Turner, B.M., 2015. Dosage compensation in mammals. *Cold spring harb perspect biol* [Online], 7(3), p.a019406.
- Brockdorff, N., 2011. Chromosome silencing mechanisms in x-chromosome inactivation: unknown unknowns. *Development*, 138(23), pp.5057–5065.
- Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., and Willard, H.F., 1992. The human xist gene: analysis of a 17 kb inactive x-specific rna that contains conserved repeats and is highly localized within the nucleus. *Cell* [Online], 71(3), pp.527–42.
- Canavez, F.C., Luche, D.D., Stothard, P., Leite, K.R., Sousa-Canavez, J.M., Plastow, G., Meidanis, J., Souza, M.A., Feijao, P., Moore, S.S., and Camara-Lopes, L.H., 2012. Genome sequence and assembly of bos indicus. *J hered* [Online], 103(3), pp.342–8.
- Carlson, B.M., 2018. *Human embryology and developmental biology e-book*. Elsevier Health Sciences.
- Carrel, L. and Willard, H.F., 2005. X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature* [Online], 434(7031), pp.400–4.

- Chao, Y., Yuan, J., Li, S., Jia, S., Han, L., and Xu, L., 2018. Analysis of transcripts and splice isoforms in red clover (*trifolium pratense* L.) by single-molecule long-read sequencing. *Bmc plant biol* [Online], 18(1), p.300.
- Chen, Z., Hagen, D.E., Wang, J., Elisk, C.G., Ji, T., Siqueira, L.G., Hansen, P.J., and Rivera, R.M., 2016. Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing. *Epigenetics* [Online], 11(7), pp.501–16.
- Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M., Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., and Schatz, M.C., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat methods* [Online], 13(12), pp.1050–1054.
- Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B., 1996. Xist rna paints the inactive x chromosome at interphase: evidence for a novel rna involved in nuclear/chromosome structure. *J cell biol* [Online], 132(3), pp.259–75.
- Collins, F.S., Morgan, M., and Patrinos, A., 2003. The human genome project: lessons from large-scale biology. *Science* [Online], 300(5617), pp.286–90.
- Cook, M.I. and Monaghan, P., 2004. Sex differences in embryo development periods and effects on avian hatching patterns. *Behavioral ecology*, 15(2), pp.205–209.
- Cotton, A.M., Price, E.M., Jones, M.J., Balaton, B.P., Kobor, M.S., and Brown, C.J., 2015. Landscape of dna methylation on the x chromosome reflects cpg density, functional chromatin state and x-chromosome inactivation. *Hum mol genet* [Online], 24(6), pp.1528–39.
- Dementyeva, E.V., Shevchenko, A.I., Anopriyenko, O.V., Mazurok, N.A., Elisaphenko, E.A., Nesterova, T.B., Brockdorff, N., and Zakian, S.M., 2010. Difference between random and imprinted x inactivation in common voles. *Chromosoma*, 119(5), pp.541–552.

- Dijk, E.L. van, Jaszczyszyn, Y., Naquin, D., and Thermes, C., 2018. The third revolution in sequencing technology. *Trends genet* [Online], 34(9), pp.666–681.
- Dindot, S.V., Kent, K.C., Evers, B., Loskutoff, N., Womack, J., and Piedrahita, J.A., 2004. Conservation of genomic imprinting at the *xist*, *igf2*, and *gtl2* loci in the bovine. *Mamm genome* [Online], 15(12), pp.966–74.
- Ding, L., Rath, E., and Bai, Y., 2017. Comparison of alternative splicing junction detection tools using rna-seq data. *Curr genomics* [Online], 18(3), pp.268–277.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R., 2013. Star: ultrafast universal rna-seq aligner. *Bioinformatics* [Online], 29(1), pp.15–21.
- Du, M., Yin, J., and Zhu, M.J., 2010. Cellular signaling pathways regulating the initial stage of adipogenesis and marbling of skeletal muscle. *Meat sci* [Online], 86(1), pp.103–9.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L., 2017. De novo assembly of the *aedes aegypti* genome using hi-c yields chromosome-length scaffolds. *Science* [Online], 356(6333), pp.92–95.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S., 2009. Real-time dna sequencing from single polymerase molecules. *Science* [Online], 323(5910), pp.133–8.

- Epstein, C.J., Travis, B., Tucker, G., and Smith, S., 1978. The direct demonstration of an x-chromosome dosage effect prior to inactivation. In: *Genetic mosaics and chimeras in mammals*. Springer, pp.261–267.
- Fang, J., Ma, Q., Chu, C., Huang, B., Li, L., Cai, P., Batista, P.J., Tolentino, K.E.M., Xu, J., Li, R., Du, P., Qu, K., and Chang, H.Y., 2019. Pirch-seq: functional classification of non-coding rnas associated with distinct histone modifications. *Genome biol* [Online], 20(1), p.292.
- Ferguson-Smith, A.C., Cattanach, B.M., Barton, S.C., Beechey, C.V., and Surani, M.A., 1991. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature* [Online], 351(6328), pp.667–70.
- Fernandes, J.C., Acuña, S.M., Aoki, J.I., Floeter-Winter, L.M., and Muxel, S.M., 2019. Long non-coding rnas in the regulation of gene expression: physiology and disease. *Non-coding rna*, 5(1), p.17.
- Frisch, J. and Vercoe, J.E., 1977. Food intake, eating rate, weight gains, metabolic rate and efficiency of feed utilization in bos taurus and bos indicus crossbred cattle. *Animal science*, 25(3), pp.343–358.
- Fujimoto, T., Nishimura, T., Goto-Kazeto, R., Kawakami, Y., Yamaha, E., and Arai, K., 2010. Sexual dimorphism of gonadal structure and gene expression in germ cell-deficient loach, a teleost fish. *Proc natl acad sci u s a* [Online], 107(40), pp.17211–6.
- Garrison, E. and Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *Arxiv preprint arxiv:1207.3907*.
- Ghurye, J., Pop, M., Koren, S., Bickhart, D., and Chin, C.S., 2017. Scaffolding of long read assemblies using long range contact information. *Bmc genomics* [Online], 18(1), p.527.
- Gilbert, W. and Maxam, A., 1973. The nucleotide sequence of the lac operator. *Proc natl acad sci u s a* [Online], 70(12), pp.3581–4.

- Gonzalez-Garay, M.L., 2016. Introduction to isoform sequencing using pacific biosciences technology (iso-seq). In: *Transcriptomics and gene regulation*. Springer, pp.141–160.
- Gordon, A. and Hannon, G., 2010. *Fastx-toolkit*. http://hannonlab.cshl.edu/fastx_toolkit.
- Grubb, B.J., 2006. Developmental biology, scott f. gilbert, editor. *Integrative and comparative biology*, 46(5), pp.652–653.
- Haberle, V. and Stark, A., 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature reviews molecular cell biology*, 19(10), pp.621–637.
- Hansen, P.J., 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim reprod sci* [Online], 82-83, pp.349–60.
- Hardcastle, T.J. and Kelly, K.A., 2010. Bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *Bmc bioinformatics* [Online], 11, p.422.
- Hartley, S.W. and Mullikin, J.C., 2016. Detection and visualization of differential splicing in rna-seq data with junctionseq. *Nucleic acids res* [Online], 44(15), e127.
- Hiremath, P.J., Kumar, A., Penmetsa, R.V., Farmer, A., Schlueter, J.A., Chamarthi, S.K., Whaley, A.M., Carrasquilla-Garcia, N., Gaur, P.M., Upadhyaya, H.D., Kavi Kishor, P.B., Shah, T.M., Cook, D.R., and Varshney, R.K., 2012. Large-scale development of cost-effective snp marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. *Plant biotechnol j* [Online], 10(6), pp.716–32.
- Hou, Y., Bickhart, D.M., Hvinden, M.L., Li, C., Song, J., Boichard, D.A., Fritz, S., Eggen, A., DeNise, S., Wiggans, G.R., Sonstegard, T.S., Van Tassell, C.P., and Liu, G.E., 2012. Fine mapping of copy number variations on two cattle genome assemblies using high density snp array. *Bmc genomics* [Online], 13, p.376.
- Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.F., Hammond, S.M., Makowski, L., Randell, S.H., Chiang, D.Y., Hayes, D.N., Jones, C., Liu, Y., Prins, J.F., and Liu, J., 2013. Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic acids res* [Online], 41(2), e39.

- Huang da, W., Sherman, B.T., and Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat protoc* [Online], 4(1), pp.44–57.
- Hughes, J.F., Skaletsky, H., Pyntikova, T., Graves, T.A., Daalen, S.K. van, Minx, P.J., Fulton, R.S., McGrath, S.D., Locke, D.P., Friedman, C., Trask, B.J., Mardis, E.R., Warren, W.C., Repping, S., Rozen, S., Wilson, R.K., and Page, D.C., 2010. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature* [Online], 463(7280), pp.536–9.
- Hwang, S., Kim, E., Lee, I., and Marcotte, E.M., 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci rep* [Online], 5, p.17875.
- Iafraite, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C., 2004. Detection of large-scale variation in the human genome. *Nat genet* [Online], 36(9), pp.949–51.
- Illumina, I., 2021. *Advantages of paired-end and single-read sequencing* [Online]. Available from: <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html> [Accessed January 6, 2021].
- Janecka, J.E., Davis, B.W., Ghosh, S., Paria, N., Das, P.J., Orlando, L., Schubert, M., Nielsen, M.K., Stout, T.A.E., Brashear, W., Li, G., Johnson, C.D., Metz, R.P., Zadjali, A.M.A., Love, C.C., Varner, D.D., Bellott, D.W., Murphy, W.J., Chowdhary, B.P., and Raudsepp, T., 2018. Horse y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nat commun* [Online], 9(1), p.2945.
- Kadri, N.K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbbrandtsen, B., Karim, L., Nielsen, U.S., Panitz, F., Aamand, G.P., Schulman, N., Georges, M., Vilkki, J., Lund, M.S., and Druet, T., 2014. A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in nordic red cattle: additional evidence for

- the common occurrence of balancing selection in livestock. *Plos genet* [Online], 10(1), e1004049.
- Kim, D., Langmead, B., and Salzberg, S.L., 2015. Hisat: a fast spliced aligner with low memory requirements. *Nat methods* [Online], 12(4), pp.357–60.
- Kim, J., Samaranyake, M., and Pradhan, S., 2009. Epigenetic mechanisms in mammals. *Cellular and molecular life sciences*, 66(4), p.596.
- Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., and Phillippy, A.M., 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat biotechnol* [Online].
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome res* [Online], 27(5), pp.722–736.
- Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, 128(4), pp.693–705.
- Kuderna, L.F.K., Lizano, E., Julia, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., Alandes, R.A., Alvarez-Estape, M., Juan, D., Simon, H., Alioto, T., Gut, M., Gut, I., Schierup, M.H., Fornas, O., and Marques-Bonet, T., 2019. Selective single molecule sequencing and assembly of a human y chromosome of african origin. *Nat commun* [Online], 10(1), p.4.
- Lahn, B.T. and Page, D.C., 1997. Functional coherence of the human y chromosome. *Science*, 278(5338), pp.675–680.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A.,

- Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al., 2001. Initial sequencing and analysis of the human genome. *Nature* [Online], 409(6822), pp.860–921.
- Law, C.W., Chen, Y., Shi, W., and Smyth, G.K., 2014. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biol* [Online], 15(2), R29.
- Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R., and Schatz, M., 2016. Third-generation sequencing and the future of genomics. *Biorxiv*, p.048603.
- Lee, T.I. and Young, R.A., 2013. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6), pp.1237–1251.
- Levine, M. and Tjian, R., 2003. Transcription regulation and animal diversity. *Nature*, 424(6945), pp.147–151.
- Liao, Y., Smyth, G.K., and Shi, W., 2013. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids res* [Online], 41(10), e108.
- Licatalosi, D.D. and Darnell, R.B., 2010. Rna processing and its regulation: global insights into biological networks. *Nature reviews genetics*, 11(1), pp.75–87.
- Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B.,

- Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., and Dekker, J., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* [Online], 326(5950), pp.289–93.
- Liu, G.E., Hou, Y., Zhu, B., Cardone, M.F., Jiang, L., Cellamare, A., Mitra, A., Alexander, L.J., Coutinho, L.L., Dell’Aquila, M.E., Gasbarre, L.C., Lacalandra, G., Li, R.W., Matukumalli, L.K., Nonneman, D., Regitano, L.C., Smith, T.P., Song, J., Sonstegard, T.S., Van Tassell, C.P., Ventura, M., Eichler, E.E., McDanel, T.G., and Keele, J.W., 2010. Analysis of copy number variations among diverse cattle breeds. *Genome res* [Online], 20(5), pp.693–703.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M., 2012. Comparison of next-generation sequencing systems. *J biomed biotechnol* [Online], 2012, p.251364.
- Liu, R., Holik, A.Z., Su, S., Jansz, N., Chen, K., Leong, H.S., Blewitt, M.E., Asselin-Labat, M.L., Smyth, G.K., and Ritchie, M.E., 2015. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic acids res* [Online], 43(15), e97.
- Liu, R., Low, W.Y., Tearle, R., Koren, S., Ghurye, J., Rhie, A., Phillippy, A.M., Rosen, B.D., Bickhart, D.M., Smith, T.P.L., Hiendleder, S., and Williams, J.L., 2019. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine x and y chromosomes. *Bmc genomics* [Online], 20(1), p.1000.
- Livesay, E. and Bee, U.G., 1945. A study of the gestation periods of five breeds of cattle. *Journal of animal science*, 4(1), pp.13–14.
- Loftus, R.T., MacHugh, D.E., Bradley, D.G., Sharp, P.M., and Cunningham, P., 1994. Evidence for two independent domestications of cattle. *Proc natl acad sci u s a* [Online], 91(7), pp.2757–61.
- Love, M.I., Huber, W., and Anders, S., 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biol* [Online], 15(12), p.550.

- Low, W.Y., Tearle, R., Bickhart, D.M., Rosen, B.D., Kingan, S.B., Swale, T., Thibaud-Nissen, F., Murphy, T.D., Young, R., Lefevre, L., Hume, D.A., Collins, A., Ajmone-Marsan, P., Smith, T.P.L., and Williams, J.L., 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat commun* [Online], 10(1), p.260.
- Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D.M., Rosen, B.D., Kronenberg, Z.N., Kingan, S.B., Tseng, E., Thibaud-Nissen, F., Martin, F.J., Billis, K., Ghurye, J., Hastie, A.R., Lee, J., Pang, A.W.C., Heaton, M.P., Phillippy, A.M., Hiendleder, S., Smith, T.P.L., and Williams, J.L., 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in angus and brahman cattle. *Nat commun* [Online], 11(1), p.2071.
- Lubchenco, L.O., Hansman, C., Dressler, M., and Boyd, E., 1963. Intrauterine growth as estimated from liveborn birth-weight data at 24 to 42 weeks of gestation. *Pediatrics* [Online], 32, pp.793–800.
- Lyon, M.F., 1962. Sex chromatin and gene action in the mammalian x-chromosome. *American journal of human genetics*, 14(2), p.135.
- MacCord, K., 2013. Germ layers. *Embryo project encyclopedia*.
- MacHugh, D.E., Larson, G., and Orlando, L., 2017. Taming the past: ancient dna and the study of animal domestication. *Annu rev anim biosci* [Online], 5, pp.329–351.
- Mak, A.C., Lai, Y.Y., Lam, E.T., Kwok, T.P., Leung, A.K., Poon, A., Mostovoy, Y., Hastie, A.R., Stedman, W., Anantharaman, T., Andrews, W., Zhou, X., Pang, A.W., Dai, H., Chu, C., Lin, C., Wu, J.J., Li, C.M., Li, J.W., Yim, A.K., Chan, S., Sibert, J., Dzakula, Z., Cao, H., Yiu, S.M., Chan, T.F., Yip, K.Y., Xiao, M., and Kwok, P.Y., 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* [Online], 202(1), pp.351–62.
- Mantere, T., Kersten, S., and Hoischen, A., 2019. Long-read sequencing emerging in medical genetics. *Frontiers in genetics*, 10, p.426.

- Martin, C. and Zhang, Y., 2005. The diverse functions of histone lysine methylation. *Nature reviews molecular cell biology*, 6(11), pp.838–849.
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet. journal*, 17(1), pp.10–12.
- Mayer, C. and Joseph, K.S., 2013. Fetal growth: a review of terms, concepts and issues relevant to obstetrics. *Ultrasound obstet gynecol* [Online], 41(2), pp.136–45.
- McKay, S.D., Schnabel, R.D., Murdoch, B.M., Matukumalli, L.K., Aerts, J., Coppieters, W., Crews, D., Dias Neto, E., Gill, C.A., Gao, C., Mannen, H., Stothard, P., Wang, Z., Van Tassell, C.P., Williams, J.L., Taylor, J.F., and Moore, S.S., 2007. Whole genome linkage disequilibrium maps in cattle. *Bmc genet* [Online], 8, p.74.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A., 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome res* [Online], 20(9), pp.1297–303.
- Medugorac, I., Seichter, D., Graf, A., Russ, I., Blum, H., Gopel, K.H., Rothhammer, S., Forster, M., and Krebs, S., 2012. Bovine polledness—an autosomal dominant trait with allelic heterogeneity. *Plos one* [Online], 7(6), e39477.
- Meyer, M. and Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold spring harb protoc* [Online], 2010(6), pdb prot5448.
- Moridi, M., Hosseini Moghaddam, S.H., Mirhoseini, S.Z., and Bionaz, M., 2019. Transcriptome analysis showed differences of two purebred cattle and their crossbreds. *Italian journal of animal science*, 18(1), pp.70–79.
- Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gilderleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., Lee, C., Turner, E.H., Smith, J.D., Rieder, M.J., Yoshiura, K.-I., Matsumoto, N., Ohta, T., Niikawa, N.,

- Nickerson, D.A., Bamshad, M.J., and Shendure, J., 2010. Exome sequencing identifies *mlh2* mutations as a cause of kabuki syndrome. *Nature genetics* [Online], 42(9) (), pp.790–793.
- Ngun, T.C., Ghahramani, N., Sanchez, F.J., Bocklandt, S., and Vilain, E., 2011. The genetics of sex differences in brain and behavior. *Front neuroendocrinol* [Online], 32(2), pp.227–46.
- NHGRI, 2020. *Chromatin structure*. <https://www.genome.gov/genetics-glossary/Chromatin>.
- Nojiri, S., Itoh, H., Kasai, T., Fujibayashi, K., Saito, T., Hiratsuka, Y., Okuzawa, A., Naito, T., Yokoyama, K., and Daida, H., 2019. Comorbidity status in hospitalized elderly in japan: analysis from national database of health insurance claims and specific health checkups. *Sci rep* [Online], 9(1), p.20237.
- O’Neil, D., Glowatz, H., and Schlumpberger, M., 2013. Ribosomal rna depletion for efficient use of rna-seq capacity. *Curr protoc mol biol* [Online], Chapter 4, Unit 4 19.
- Oestrup, O., Hall, V., Petkov, S.G., Wolf, X.A., Hyldig, S., and Hyttel, P., 2009. From zygote to implantation: morphological and molecular dynamics during embryo development in the pig. *Reprod domest anim* [Online], 44 Suppl 3, pp.39–49.
- Oliveto, S., Mancino, M., Manfrini, N., and Biffo, S., 2017. Role of micrnas in translation regulation and cancer. *World journal of biological chemistry*, 8(1), p.45.
- Otis, E.M. and Brent, R., 1954. Equivalent ages in mouse and human embryos. *Anat rec* [Online], 120(1), pp.33–63.
- Pal, K., Forcato, M., and Ferrari, F., 2019. Hi-c analysis: from data generation to integration. *Biophysical reviews*, 11(1), pp.67–78.
- Palazzo, A.F. and Lee, E.S., 2015. Non-coding rna: what is functional and what is junk? *Frontiers in genetics*, 6, p.2.

- Pareek, C.S., Sachajko, M., Jaskowski, J.M., Herudzinska, M., Skowronski, M., Domagalski, K., Szczepanek, J., Czarnik, U., Sobiech, P., Wysocka, D., Pierzchala, M., Polawska, E., Stepanow, K., Ogluszka, M., Juszczuk-Kubiak, E., Feng, Y., and Kumar, D., 2019. Comparative analysis of the liver transcriptome among cattle breeds using rna-seq. *Vet sci* [Online], 6(2).
- Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat methods* [Online], 14(4), pp.417–419.
- Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A.K., Pandey, A., and Salzberg, S., 2018. Thousands of large-scale rna sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *Biorxiv*, p.332825.
- Pitt, D., Sevane, N., Nicolazzi, E.L., MacHugh, D.E., Park, S.D.E., Colli, L., Martinez, R., Bruford, M.W., and Orozco-terWengel, P., 2019. Domestication of cattle: two or three events? *Evol appl* [Online], 12(1), pp.123–136.
- Qin, Q., Mei, S., Wu, Q., Sun, H., Li, L., Taing, L., Chen, S., Li, F., Liu, T., Zang, C., Xu, H., Chen, Y., Meyer, C.A., Zhang, Y., Brown, M., Long, H.W., and Liu, X.S., 2016. Chilin: a comprehensive chip-seq and dnase-seq quality control and analysis pipeline. *Bmc bioinformatics* [Online], 17(1), p.404.
- Regitz-Zagrosek, V., 2012. Sex and gender differences in health. *Embo reports*, 13(7), pp.596–603.
- Robert, C. and Watson, M., 2015. Errors in rna-seq quantification affect genes of relevance to human disease. *Genome biol* [Online], 16, p.177.
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K., 2010. Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* [Online], 26(1), pp.139–40.

- Rozen, S., Warren, W.C., Weinstock, G., and O'Brien, S.J., 2006. Sequencing and annotating new mammalian y chromosomes.
- Sandmann, S., Graaf, A.O. de, Karimi, M., Reijden, B.A. van der, Hellstrom-Lindberg, E., Jansen, J.H., and Dugas, M., 2017. Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci rep* [Online], 7, p.43169.
- Sanger, F., Nicklen, S., and Coulson, A.R., 1977. Dna sequencing with chain-terminating inhibitors. *Proc natl acad sci u s a* [Online], 74(12), pp.5463–7.
- Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* [Online], 270(5235), pp.467–70.
- Schibler, J. and Schlumbaum, A., 2007. [history and economic importance of cattle (bos taurus l.) in switzerland from neolithic to early middle ages]. *Schweiz arch tierheilkd* [Online], 149(1), pp.23–9.
- Schmidt-Rhaesa, A., 2007. *The evolution of organ systems*. Oxford University Press.
- Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., et al., 2016. How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *Rna*, 22(6), pp.839–851.
- Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y., 2014. Rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data. *Proc natl acad sci u s a* [Online], 111(51), E5593–601.
- Shi, Y. and Jiang, H., 2013. Rseqdiff: detecting differential isoform expression from rna-seq data using hierarchical likelihood ratio test. *Plos one* [Online], 8(11), e79448.
- Shlyueva, D., Stampfel, G., and Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews genetics*, 15(4), pp.272–286.

- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.F., Latrielle, P., Leonard, S., Mardis, E., Maupin, R., McPherson, J., Miner, T., Nash, W., Nguyen, C., Ozersky, P., Pepin, K., Rock, S., Rohlfsing, T., Scott, K., Schultz, B., Strong, C., Tin-Wollam, A., Yang, S.P., Waterston, R.H., Wilson, R.K., Rozen, S., and Page, D.C., 2003. The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *Nature* [Online], 423(6942), pp.825–37.
- Skinner, B.M., Sargent, C.A., Churcher, C., Hunt, T., Herrero, J., Loveland, J.E., Dunn, M., Louzada, S., Fu, B., Chow, W., Gilbert, J., Austin-Guest, S., Beal, K., Carvalho-Silva, D., Cheng, W., Gordon, D., Grafham, D., Hardy, M., Harley, J., Hauser, H., Howden, P., Howe, K., Lachani, K., Ellis, P.J., Kelly, D., Kerry, G., Kerwin, J., Ng, B.L., Threadgold, G., Wileman, T., Wood, J.M., Yang, F., Harrow, J., Affara, N.A., and Tyler-Smith, C., 2016. The pig x and y chromosomes: structure, sequence, and evolution. *Genome res* [Online], 26(1), pp.130–9.
- Snelling, W.M., Chiu, R., Schein, J.E., Hobbs, M., Abbey, C.A., Adelson, D.L., Aerts, J., Bennett, G.L., Bosdet, I.E., Boussaha, M., Brauning, R., Caetano, A.R., Costa, M.M., Crawford, A.M., Dalrymple, B.P., Eggen, A., Everts-van der Wind, A., Floriot, S., Gautier, M., Gill, C.A., Green, R.D., Holt, R., Jann, O., Jones, S.J., Kappes, S.M., Keele, J.W., Jong, P.J. de, Larkin, D.M., Lewin, H.A., McEwan, J.C., McKay, S., Marra, M.A., Mathewson, C.A., Matukumalli, L.K., Moore, S.S., Murdoch, B., Nicholas, F.W., Osoegawa, K., Roy, A., Salih, H., Schibler, L., Schnabel, R.D., Silveri, L., Skow, L.C., Smith, T.P., Sonstegard, T.S., Taylor, J.F., Tellam, R., Van Tassell, C.P., Williams, J.L., Womack, J.E., Wye, N.H., Yang, G., Zhao, S., and International Bovine, B.A.C.M.C., 2007. A physical map of the bovine genome. *Genome biol* [Online], 8(8), R165.
- Soh, Y.Q., Alfoldi, J., Pyntikova, T., Brown, L.G., Graves, T., Minx, P.J., Fulton, R.S., Kremitzki, C., Koutseva, N., Mueller, J.L., Rozen, S., Hughes, J.F., Owens, E., Womack, J.E., Murphy, W.J., Cao, Q., Jong, P. de, Warren, W.C., Wilson, R.K., Skaletsky, H., and Page, D.C., 2014. Sequencing the mouse y chromosome reveals convergent gene

- acquisition and amplification on both sex chromosomes. *Cell* [Online], 159(4), pp.800–13.
- Steijger, T., Abril, J.F., Engstrom, P.G., Kokocinski, F., Consortium, R., Hubbard, T.J., Guigo, R., Harrow, J., and Bertone, P., 2013. Assessment of transcript reconstruction methods for rna-seq. *Nat methods* [Online], 10(12), pp.1177–84.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc natl acad sci u s a* [Online], 102(43), pp.15545–50.
- Thurston, A., Taylor, J., Gardner, J., Sinclair, K.D., and Young, L.E., 2008. Monoallelic expression of nine imprinted genes in the sheep embryo occurs after the blastocyst stage. *Reproduction* [Online], 135(1), pp.29–40.
- Todeschini, A.-L., Georges, A., and Veitia, R.A., 2014. Transcription factors: specific dna binding and specific gene regulation. *Trends in genetics*, 30(6), pp.211–219.
- Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B.B., Castel, S.E., Karczewski, K.J., Aguet, F., Byrnes, A., Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E.B.I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U.o.C.S.C., Lapalainen, T., Regev, A., Ardlie, K.G., Hacohen, N., and MacArthur, D.G., 2017. Landscape of x chromosome inactivation across human tissues. *Nature* [Online], 550(7675), pp.244–248.

- Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S., 2008. Snp discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat methods* [Online], 5(3), pp.247–52.
- Wang, J., Duncan, D., Shi, Z., and Zhang, B., 2013. Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic acids res* [Online], 41(Web Server issue), W77–83.
- Watson, J.D. and Cook-Deegan, R.M., 1991. Origins of the human genome project. *Faseb j* [Online], 5(1), pp.8–11.
- Watson, J. and Crick, F.H., 1953. Molecular structure of nucleic acids. *Nature*, 171(4356), pp.737–738.
- Weinhold, B., 2006. Epigenetics: the science of change. *Environmental health perspectives* [Online], 114(3) (), A160–A167.
- Wu, D. and Smyth, G.K., 2012. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids res* [Online], 40(17), e133.
- Wu, R. and Kaiser, A.D., 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda dna. *J mol biol* [Online], 35(3), pp.523–37.
- Xiang, R., Ghanipoor-Samami, M., Johns, W.H., Eindorf, T., Rutley, D.L., Kruk, Z.A., Fitzsimmons, C.J., Thomsen, D.A., Roberts, C.T., Burns, B.M., Anderson, G.I., Greenwood, P.L., and Hiendleder, S., 2013. Maternal and paternal genomes differentially affect myofibre characteristics and muscle weights of bovine fetuses at midgestation. *Plos one* [Online], 8(1), e53402.
- Xiang, R., Lee, A.M., Eindorf, T., Javadmanesh, A., Ghanipoor-Samami, M., Gugger, M., Fitzsimmons, C.J., Kruk, Z.A., Pitchford, W.S., Leviton, A.J., Thomsen, D.A., Beckman, I., Anderson, G.I., Burns, B.M., Rutley, D.L., Xian, C.J., and Hiendleder, S., 2014. Widespread differential maternal and paternal genome effects on fetal bone phenotype at mid-gestation. *J bone miner res* [Online], 29(11), pp.2392–404.

- Zeder, M.A., 2015. Core questions in domestication research. *Proc natl acad sci u s a* [Online], 112(11), pp.3191–8.
- Zemel, S., Bartolomei, M.S., and Tilghman, S.M., 1992. Physical linkage of two mammalian imprinted genes, h19 and insulin-like growth factor 2. *Nat genet* [Online], 2(1), pp.61–5.
- Zeng, J.Y., Robertson, I.D., Ji, Q.M., Dawa, Y.L., and Bruce, M., 2019. Evaluation of the economic impact of brucellosis in domestic yaks of tibet. *Transbound emerg dis* [Online], 66(1), pp.476–487.
- Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and Schack, D. von, 2018. Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rna depletion. *Sci rep* [Online], 8(1), p.4781.
- Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marcais, G., Roberts, M., Subramanian, P., Yorke, J.A., and Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, *bos taurus*. *Genome biol* [Online], 10(4), R42.

2 New Insights into Mammalian Sex Chromosome Structure and Evolution using High-Quality Sequences from Bovine X and Y Chromosomes

Ruijie Liu¹, Wai Yee Low¹, Rick Tearle¹, Sergey Koren², Jay Ghurye³, Arang Rhie², Adam M. Phillippy², Benjamin D. Rosen⁴, Derek M. Bickhart⁵, Timothy P.L. Smith⁶, Stefan Hiendleder¹, John L. Williams¹

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia

²Genomic Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

³Center for Bioinformatics and Computational Biology, Lab 3104A, Biomolecular Science Building, University of Maryland, College Park, Maryland, USA

⁴Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, Maryland, USA

⁵Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, Wisconsin, USA

⁶US Meat Animal Research Centre, ARS USDA, Clay Centre, Nebraska, USA

Published in 2019, BMC Genomics 20 (1), 1-11

My contribution to this publication was to validate the sex chromosome assemblies using bovine markers then to annotate the sex chromosomes and compared them to other mammalian sex chromosome assemblies. I designed figures and tables and drafted the manuscript.


Contributions of the co-authors were to provide tissue and DNA samples which were sequenced, and to curate the sequencing and Hi-C data; assemble the contigs and provide guidance on scaffolding; consolidate all data. I then collaborated with my colleagues to produce the final chromosome-level genome assembly (see the supporting publication in section § 6 of this thesis).

RESEARCH ARTICLE

Open Access

New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes



Ruijie Liu¹, Wai Yee Low¹, Rick Tearle¹, Sergey Koren², Jay Ghurye³, Arang Rhie², Adam M. Phillippy², Benjamin D. Rosen⁴, Derek M. Bickhart⁵, Timothy P. L. Smith⁶, Stefan Hiendleder¹ and John L. Williams^{1*} 

Abstract

Background: Mammalian X chromosomes are mainly euchromatic with a similar size and structure among species whereas Y chromosomes are smaller, have undergone substantial evolutionary changes and accumulated male specific genes and genes involved in sex determination. The pseudoautosomal region (PAR) is conserved on the X and Y and pair during meiosis. The structure, evolution and function of mammalian sex chromosomes, particularly the Y chromosome, is still poorly understood because few species have high quality sex chromosome assemblies.

Results: Here we report the first bovine sex chromosome assemblies that include the complete PAR spanning 6.84 Mb and three Y chromosome X-degenerate (X-d) regions. The PAR comprises 31 genes, including genes that are missing from the X chromosome in current cattle, sheep and goat reference genomes. Twenty-nine PAR genes are single-copy genes and two are multi-copy gene families, OBP, which has 3 copies and BDA20, which has 4 copies. The Y chromosome X-d1, 2a and 2b regions contain 11, 2 and 2 gametologs, respectively.

Conclusions: The ruminant PAR comprises 31 genes and is similar to the PAR of pig and dog but extends further than those of human and horse. Differences in the pseudoautosomal boundaries are consistent with evolutionary divergence times. A bovidae-specific expansion of members of the lipocalin gene family in the PAR reported here, may affect immune-modulation and anti-inflammatory responses in ruminants. Comparison of the X-d regions of Y chromosomes across species revealed that five of the X-Y gametologs, which are known to be global regulators of gene activity and candidate sexual dimorphism genes, are conserved.

Keywords: Genomes, Livestock, Bovine, Sex chromosomes, Pseudoautosomal region

Background

The sex chromosomes evolved from ancestral autosomes in dioecious lineages and have become extensively differentiated in structure and gene content [1, 2]. Mammalian X chromosomes are mainly euchromatic with a similar size and structure among species, and have retained most of the ancestral X genes [3, 4]. In contrast, Y chromosomes have undergone substantial evolutionary changes, accumulated male specific genes and genes involved in sex determination,

and have lost 95% of the ancestral genes [5]. As a consequence, the Y chromosome is much smaller than the X chromosome and comprises mainly the pseudoautosomal (PAR), X-degenerate (X-d) and ampliconic regions [6].

The PAR is conserved on the X and Y, pairing and recombining at meiosis [7]. Most mammals have a single PAR region but the human sex chromosomes are an exception with a second PAR at the distal ends of the X and Y chromosomes [8]. The PAR plays an essential role in normal sexual development and loss of the PAR is associated with male sterility in humans [9]. Despite its critical role in fertility and disease, the PAR is one of the least well-characterised parts of most mammalian

* Correspondence: john.williams01@adelaide.edu.au

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia
Full list of author information is available at the end of the article



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

genomes. Previous studies have described the genes at the bovine pseudoautosomal boundary and PAR gene content [10–12], but currently there is neither a complete gene map nor a precise size available for the bovine PAR.

The X-d regions of the Y chromosome contain single-copy genes, pseudogenes, which appear to be surviving relics of the autosomes from which the Y chromosome evolved, and ampliconic regions, which consist of large heterochromatic blocks rich in repetitive sequences [9, 13, 14]. Both the X-d and ampliconic regions are male-specific. Unlike the highly conserved PAR, the structure and gene content of the X-d regions differ among mammalian species. The human X-d regions are interrupted by several large blocks of ampliconic sequences, while the X-d regions of chimpanzee include a single ampliconic block [9, 15]. A study of bovine Y chromosome gene expression has contributed information on genes in X-d regions [16], but interpretation of the data is limited by the relatively poor quality of the available Y chromosome assembly.

The complex and highly repetitive Y ampliconic regions are difficult to assemble, particularly from short sequence reads. Many mammalian genome sequencing projects have used a female subject to avoid having to resolve X and Y haplotypes, and therefore do not include the Y chromosome. Only a few species, including human [6], chimpanzee [15], rhesus macaque [17], mouse [18], pig [19] and horse [20], have well characterised and assembled Y chromosomes.

The *Bos taurus taurus* reference genome assembly Btau_5.0.1 [21] (NCBI Project ID:20275) was assembled from short and long sequence reads of BAC clones and contains a Y chromosome sequence. A *Bos taurus indicus* Y chromosome was created by alignment of short read sequences to Btau_4.0 [22], and therefore will be missing any larger indicine-specific features. These sex chromosome assemblies are incomplete and inconsistent, hindering studies on sex chromosome evolution and the dissection of the molecular architecture of sexually dimorphic phenotypic traits.

In the present article, we report high quality assemblies of bovine X and Y chromosomes, created from long read sequences and optical mapping data, using a trio binning approach [23] that exploited the high level of DNA sequence divergence between the two subspecies of domestic cattle [24–26], *Bos taurus taurus* and *Bos taurus indicus*. We present a detailed gene map of the complete bovine PAR and X-d regions from these assemblies and discuss the evolutionary changes and functional aspects in these regions in comparison with other mammals.

Results

Assembly and annotation of the cattle sex chromosomes

The bovine X and Y chromosomes were assembled from whole genome sequence of a hybrid male with a

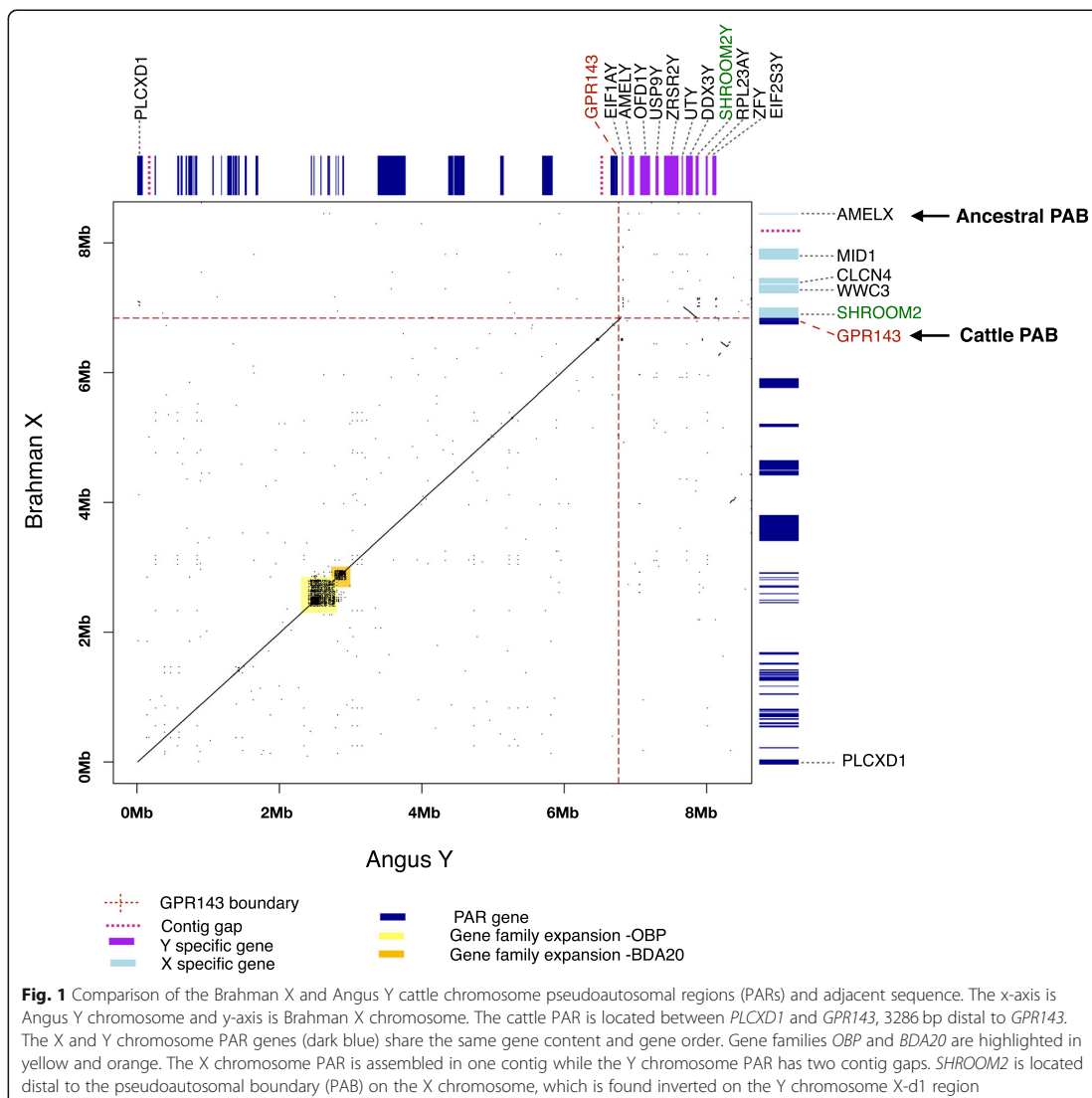
Bos taurus taurus (Angus) sire and a *Bos taurus indicus* (Brahman) dam [23] (see URLs). The assembled chromosomes presented here are the Brahman X chromosome which comprises 146 Mb in 106 contigs with 983 genes, and the Angus Y chromosome which comprises 16 Mb in 67 contigs with 51 unique genes (Table 1 and Additional file 1: Table S1). These sequence assemblies have been deposited at NCBI (X: CM0011833.1; Y: CM0011803.1). The full length of the cattle Y chromosome has been estimated as ~ 50 Mb, at least half of which is in the highly repetitive region [27]. As in other species [19, 20, 28], even with long read sequencing, we could not assemble the ampliconic highly repetitive region [27] or the heterochromatic regions. Full annotation of the Brahman X and Angus Y chromosomes are available from Ensembl release v97 (UOA_Brahaman_1 and UOA_Angus_1). Analysis of the PAR and X-degenerate regions are presented below.

Identification of the cattle PAR

Alignment of assembled Brahman X and Angus Y chromosomes to each other identified a 6.8 Mb region with 99% sequence identity that extends from the start of the assembled X chromosome sequence (CM0011833.1) to 2933 bp distal to *GPR143*, after which sequence identity decreases to 86% for 348 bp and then drops abruptly to ~ 15% for the next 1 Mb (Fig. 1). The X chromosome PAR is assembled in one contig while the Y chromosome PAR only has two contig gaps. This enabled us to precisely define the PAR boundary and size. The PAR on the Brahman X and Angus Y chromosomes contained 31 genes in the same order. Of these, 29 are single-copy genes and two are multi-copy gene families, *OBP*, which has 3 copies and *BDA20*, which has 4 copies (Additional file 1: Table S4). The Brahman X chromosome PAR contains 12 genes that are missing from the proximal end of the X

Table 1 Length and number of gaps for Mammalian Sex chromosomes

	X Length (bp)	X Gaps	Y Length (bp)	Y Gaps
Cattle-Brahman/Angus	146,049,346	91	15,624,455	69
Cattle-Nelore	82,205,613	10,873	14,991,264	35,040
Cattle-Hereford	139,002,886	55	38,719,986	18
Water buffalo	143,477,029	65	–	–
Goat	115,943,529	319	–	–
Sheep	132,936,813	2968	–	–
Pig	125,778,992	10	15,567,420	12
Dog	123,180,702	1032	–	–
Horse	127,806,490	300	8,967,074	560
Human	154,893,106	28	26,415,094	55

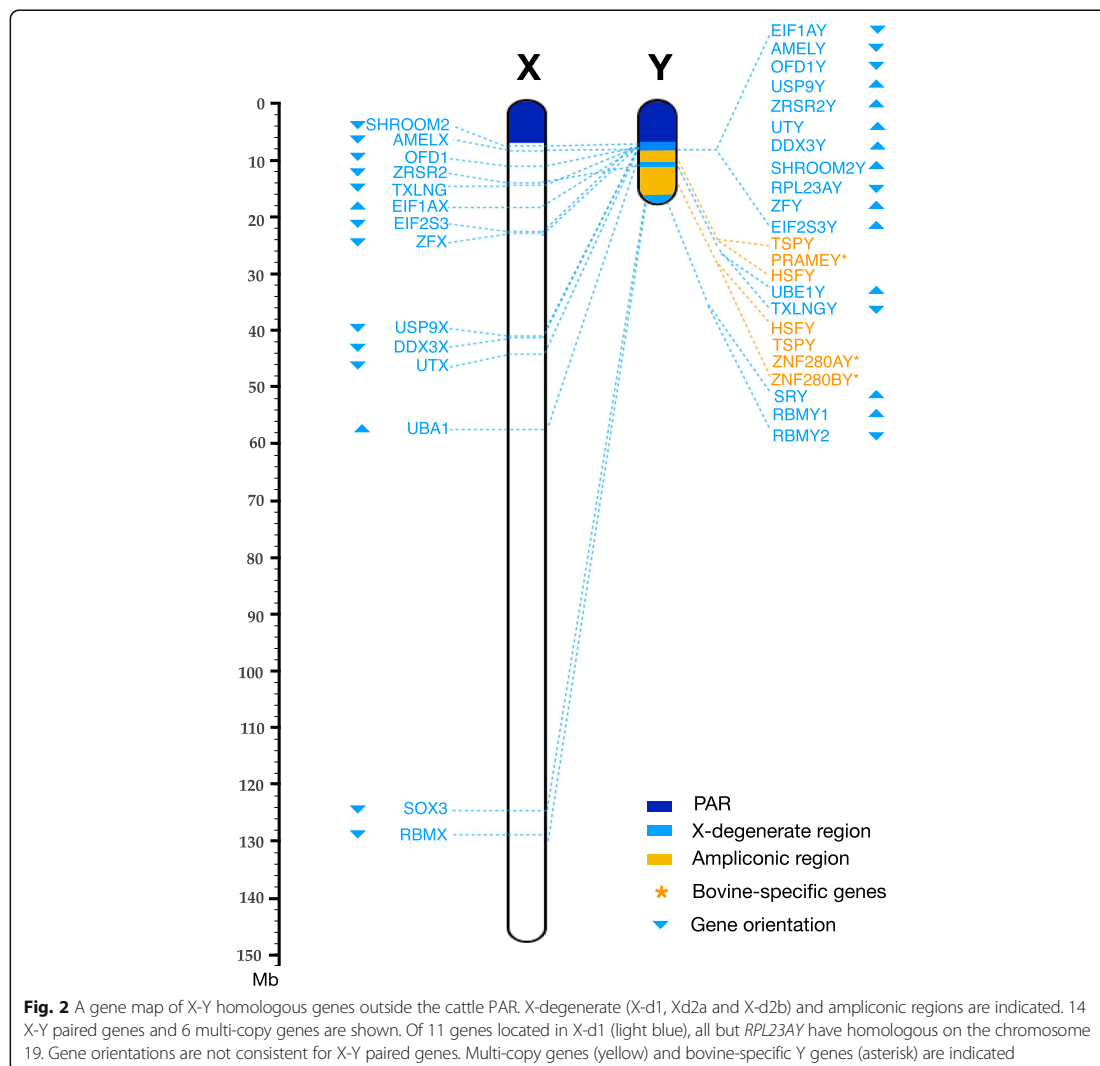


chromosome in the current Hereford reference genome ARS-UCD1.2 (Additional file 1: Table S3).

Identification of cattle X-degenerate regions

Additional genes outside the PAR showed between 60 and 96% sequence identity between the X and Y chromosomes and are located in X-degenerate regions of the Y chromosome. The first of these regions, X-d1, is located distal to the PAR and spans 1.48 Mb, between 6.84 Mb and 8.32 Mb. X-d1 contains 11 single-copy protein coding genes. The corresponding region on the X chromosome spans 35 Mb. and contains 10 X-d1

homologues in a different order but misses *RPL23AY*, which is located on chromosome 19 (Fig. 2). A 3 Mb ampliconic region immediately distal to X-d1 contains the male-specific Y (MSY) gene families *PRAMEY*, *TSPY*, and *HSFY*. At the distal end of the ampliconic region, the second X-degenerate region, X-d2a, spans 1.63 Mb and contains two single copy genes, *UBE1Y* and *TXLNGY*. The X chromosome homologs of these two genes are separated by a 44 Mb interval that contains 285 X chromosome-specific genes. Distal to X-d2a lies a 4.5 Mb ampliconic segment containing the bovine specific MSY genes *ZNF280AY* and *ZNF280BY*, which are



equivalent to *TSPY* and *HSFY* found in other species. The copy numbers of multi-copy MSY gene families are listed in Additional file 1: Table S2 and the complex arrangement of multi-copy genes is presented in Additional file 1: Figure S3. The distal end of chromosome Y contains the third X-degenerate region, X-d2b, which extends over 1.3 Mb and includes *SRY* and two copies of *RBMY*. The X chromosome homologs of these, *SOX3* and *RBMX*, are located in a 5 Mb segment at the distal end of the X chromosome.

Comparison of sex chromosome structure in mammals

Alignment of the Brahman X chromosome with the current *Bos taurus taurus* (Hereford) cattle reference

sequence (ARS-UCD1.2) revealed a 4 Mb inversion as a major structural difference. In both assemblies this inverted region ends at contig breakpoints (Additional file 1: Figure S1a). Alignment of the Brahman X chromosome with the water buffalo X chromosome [29] revealed a high level of co-linearity, with one large inversion and five small inversions at the distal end of the chromosome (Additional file 1: Figure S1b). The Brahman and water buffalo X chromosomes are 30 and 25 Mb longer, respectively, than the goat X chromosome, which consists of two scaffolds with a combined length of 116 Mb [30]. The goat X chromosome shows excellent co-linearity overall with the sheep X chromosome (Additional file 1: Figure S1c-d) but both showed numerous break points

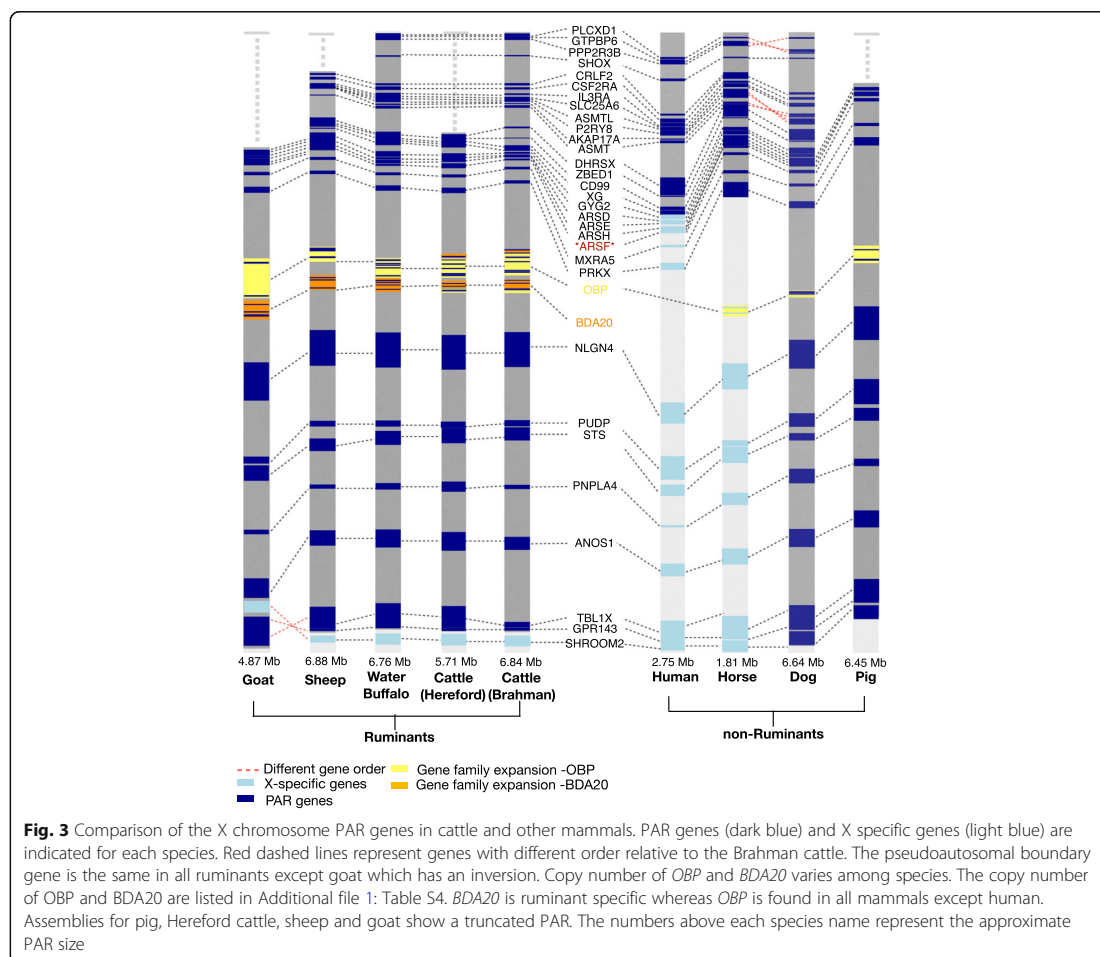
and several inversions, particularly on the short arm, in comparison with the Brahman and water buffalo X chromosomes. Non-ruminant mammalian X chromosomes, i.e. human, pig, dog and horse, revealed a striking similarity in the pattern of rearrangements in comparison to the *Bos taurus indicus* (Brahman) X chromosome (Additional file 1: Figure S1e-h). These consisted predominantly of 5 large inversions.

Alignment of the Angus Y chromosome assembly with pig, horse and human Y chromosomes showed limited co-linearity which was confined to the PAR and X-degenerate regions (Additional file 1: Figure S2a-c).

Gene content and order of the mammalian PAR

There is a very high level of conservation of synteny among mammalian PARs (Fig. 3). *PLCXD1* is the most proximal PAR gene in human, horse, Brahman cattle and water buffalo. At their proximal ends, the

PAR regions in the Hereford cattle reference genome, and sheep, goat and pig assemblies are truncated distal to *DHRXS*, *CLRF2*, *CD99* and *GYG2* respectively (Additional file 1: Table S3). At their distal end, the pig and dog PAR extend beyond *GPR143* with a boundary distal to *SHROOM2*. In comparison to all the other species, the goat PAR has an inversion of three genes (*TBLIX*, *GPR143*, *SHROOM2*) close to the ruminant PAR boundary. This region is contained in one contig of the goat assembly and may thus be a contig orientation error, rather than a goat-specific rearrangement. The human sex chromosomes are an exception amongst mammals and have PARs at the proximal and distal ends [8]. The PAR1 in human is equivalent to the single PAR of other mammalian species, but is much shorter, with a distal boundary proximal to *XG*. The PAR of horse is the shortest with the distal boundary at *PRKX* (Fig. 3).



PAR gene family expansions in different lineages

The *OBP* gene family, which is distal to *PRKX* in all species, is within the PAR of all ruminants, pig and dog, but is outside the PAR of horse, and is missing from the human X chromosome. This gene family is expanded in ruminants (Fig. 3). The *BDA20* gene family is immediately distal to the *OBP* family and present in all ruminants for which data are available, including Yak [31], Deer [32] and Chiru [33], but is not found in other mammals (Fig. 3). The *BDA20* family shows differential expansion in the different ruminant species, with two or more copies with 74–91% nucleotide sequence identity at mRNA level in cattle [21], sheep [34], goat [30] and buffalo [29] (Fig. 3, Additional file 1: Table S4). In contrast, *ARSF*, a member of the *ARS* family, has been reported as a PAR gene in other mammalian species, but is not found in any of the ruminant PARs [29, 30, 34].

Comparison of X-degenerate Y chromosome regions

Most of the X-Y paired genes of cattle, pig and horse that are outside the PAR are found in the X degenerate region, X-d1, located adjacent to the PAR (Fig. 4). Of the 11 genes in the cattle X-d1 region 8 are in common with horse and pig X-d1 regions, but the gene order differs between the three species. *RPL23AY* is only found

in the cow X-d1, while *TMSB4Y* is found in the horse and pig X-d1 regions and the human X-d3 region which is missing from cow X-d regions. Five additional bovine gametologs are found in two X-d2 regions, X-d2a and X-d2b, which correspond to the single X-d2 in horse and pig. Cattle X-d2a is distal to X-d1 and contains 2 genes, *UBE1Y* and *TXLNGY*. Both genes are found in the pig X-d2 region but *UBE1Y* is in an ampliconic region of the horse Y chromosome. The cattle X-d2b region contains *SRY* and is in a telomeric position similar to the X-d2 region of pig. The cattle X-d2b region contains two copies of *RBMY*, which is also duplicated in the horse X-d2 [20].

Discussion

The trio-binning approach facilitated the construction of the most complete bovine X and Y chromosome sequence assemblies available to date. Alignment of the Brahman cattle X with Hereford X revealed a major inverted region. In both assemblies this inverted region ends at contig breakpoints and could be an assembly artefact rather than a true biological difference. We could not resolve this discrepancy using RH and linkage maps [35–37] because the marker density was insufficient. Alignments of the Brahman cattle X chromosome

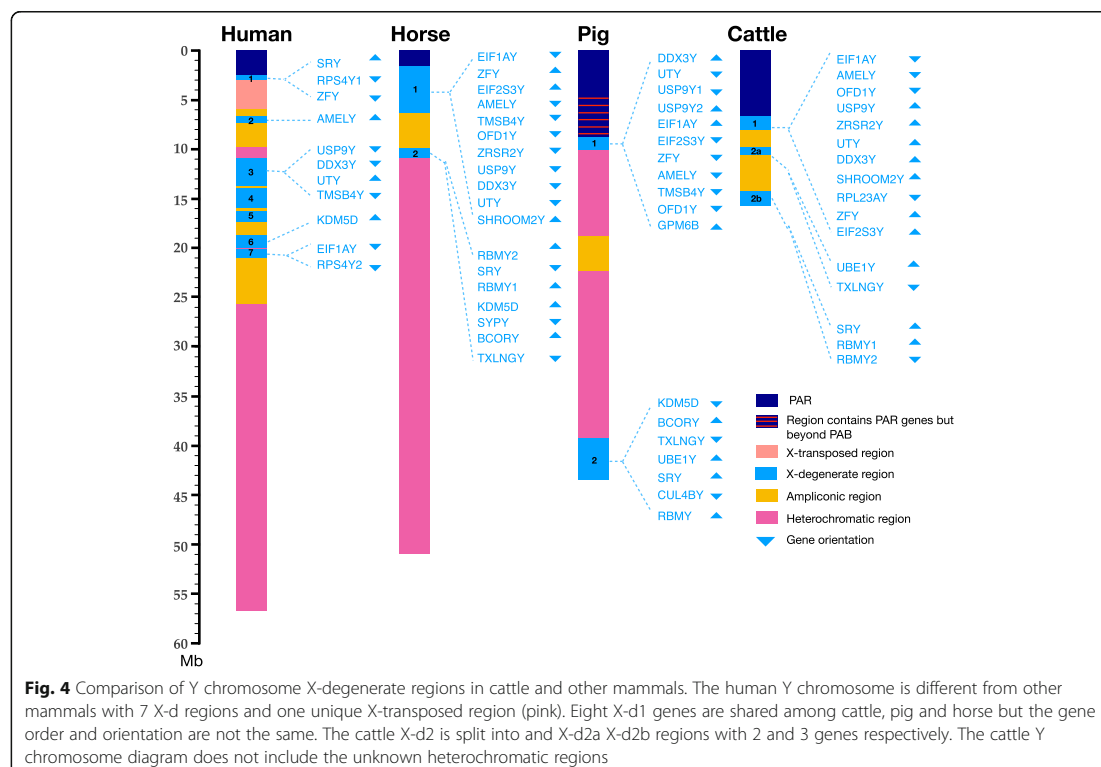


Fig. 4 Comparison of Y chromosome X-degenerate regions in cattle and other mammals. The human Y chromosome is different from other mammals with 7 X-d regions and one unique X-transposed region (pink). Eight X-d1 genes are shared among cattle, pig and horse but the gene order and orientation are not the same. The cattle X-d2 is split into and X-d2a X-d2b regions with 2 and 3 genes respectively. The cattle Y chromosome diagram does not include the unknown heterochromatic regions

with those of the water buffalo, goat and sheep, revealed a high level of co-linearity, but with numerous break points and several inversions. Many of these structural differences are consistent with gross karyotypic rearrangements that differentiate goat and sheep X chromosomes from those of cattle and water buffalo [38]. The evolutionary separation of goat and sheep from cattle and buffalo occurred 20–25 MYBP, which pre-dates the separation of goat and sheep from each other, 10–15 MYBP, and of buffalo from cattle, about 12 MYBP [39]. This evolutionary history is consistent with the differences in X chromosome structure we see among these species [24, 40].

The X chromosomes of human, pig, dog and horse have a strikingly similar pattern of differences relative to the Brahman cattle X chromosome. Ruminants are even-toed ungulates which separated from odd-toed ungulates (including the pig and horse) more than 60 MYBP, suggesting that the rearrangements occurred after this split but prior to the separation of the ruminants, about 25 MYBP [24].

Comparison of the Brahman cattle X chromosome with other mammalian reference genomes suggest the X chromosome assemblies for the Hereford cattle, sheep, goat and pig are incomplete at the proximal end of the PAR. The proximal PAR genes of Brahman cattle, water buffalo, human and horse are the same. Genes missing from the proximal end of the Hereford cattle X chromosome have been misplaced on various autosomes, whereas the genes missing from sheep, goat and pig, are found on unplaced contigs. Given that the common ancestor of cattle, water buffalo, human and horse existed about 96 MYBP, it is more parsimonious to suggest the gene order and structure of the proximal PAR are conserved and not assembly artefacts. The assembly of the Brahman cattle and water buffalo [29] X chromosomes may have benefited from the use of much longer PacBio reads and improved assembly algorithms.

The distal PAR boundary in ancient species is thought to lie within the *AMEL* locus [41]. However, there was no substantial identity of the Brahman X and Angus Y chromosomes for the region between 2933 bp distal to *GPR143* and *AMELX*. The PAR boundary in cattle is therefore just distal to *GPR143*. The distal ends of the PARs of water buffalo and sheep also lie close to *GPR143*.

The expansion of *OBP* and *BDA20* gene families in ruminants suggests they have a specific role in these species. These genes are members of the lipocalin family which are involved in immune-modulation and anti-inflammatory responses [42]. For example, their expression changes after exposure of cattle to ticks [43]. The *ARS* gene family in the ruminant PAR is missing *ARSF* that is found in non-ruminant species, suggesting that it has been lost during ruminant evolution.

Recombination of mammalian sex chromosomes only occurs in the PAR. This may explain why the MSY regions are rich in repetitive sequences [2, 44]. The co-linearity between Angus and Hereford Y chromosomes is limited to X-d and *PRAMEY* regions, while the numerous repetitive sequences in ampliconic regions appear expanded in the Hereford Y. This may be due to the use of various BAC clones from several individuals to assemble the Hereford Y chromosome [45]. The Angus Y chromosome assembly shows some alignment in some isolated areas with pig, horse and human Y chromosomes, which is mainly in X-degenerate and ampliconic regions. This is consistent with rapid evolution of non-recombining Y chromosome sequences [3].

We identified 16 X-Y paired genes in 3 X-degenerate regions on the Angus Y chromosome, which were interspersed by ampliconic regions that contain multiple copies of bovine-specific *PRAMEY*, *ZNF280AY* and *ZNF280BY* genes. *PRAMEY* genes are exclusively expressed in testis and are involved in spermatogenesis during testicular maturation [46]. *ZNF280BY* and *ZNF280AY* are multi-copy Y-genes transposed from an autosome. The temporal and spatial expression patterns of these genes also suggests that they play a role in spermatogenesis [47]. While horse, pig and cattle X-d1 are similar in gene content, gene order is very different in the three species. *RPL23AY* is cattle-specific and a second copy of this gene, *RPL23A*, is found on chromosome 19, located in a conserved block in mammals with intron-exon structure and six identical intron-less pseudogenes with 91% sequence identity to this gene are found on cattle chromosomes 3, 9, 10, 14, 22 and 29, suggesting that these copies have arisen by retrotransposition.

Cattle have lost several ancestral Y genes from X-d regions including *KD5MD*, *TMSB4Y* and *TXLNGY*. Ten of the 16 gametologs are conserved in the X-d regions of horse, 8 in pig and 5 in human. The genes in the X-d regions, do not recombine, and hence diverge over time, allowing for the possibility of sex-specific selection. They may therefore be involved in sexual dimorphism [48]. Five of these (*DDX3Y*, *EIF1AY*, *USP9Y*, *UTY*, *ZFY*) are global regulators of gene activity expressed across a broad range of adult tissues and could have profound effects on sexual development [48].

Conclusions

The quality of assemblies achieved for the X chromosome, and for the PAR, X-d and ampliconic regions of the Y chromosome, using the trio-binning approach, enabled us to examine and compare major structures of the bovine sex chromosomes, both between the sex

chromosomes and among species. Alignment of the Brahman X and Angus Y chromosomes precisely identified boundaries and gene content of the PAR region and indicated that the proximal end of the PAR is truncated in the sheep, goat and current bovine reference genome assemblies. The sequence data revealed expansions of gene families in the ruminant PAR region that have previously been associated with immune function, and conservation of gametologs that are known dosage sensitive regulators of gene expression. The sex chromosome assemblies and the annotation presented are valuable resources for the molecular characterization of sex-specific phenotype in livestock and other species.

Methods

Sample collection

All animal work was approved by the University of Adelaide Animal Ethics Committee (No. S-094-2005). Briefly, a Brahman cow was bought by the University of Adelaide from a farm (Kiowa, Kingstown, New South Wales) and was transported to SARDI experimental farm at Struan South Australia where it was inseminated with semen of an Angus bull bought by the University from American Breeder Services, Australia. At day 153 post-insemination, the cow was humanely killed by stunning and exsanguination at a commercial abattoir (Dalriada Abattoir, Keith, South Australia) as per standard operating procedures. The uterus was recovered and the male fetus removed and immediately humanely killed by stunning and exsanguination. Lung tissue was collected and snap frozen in liquid nitrogen. Details of the contig creation for this assembly using the trio binning method have been previously described [23]. Briefly, DNA was extracted from fetal lung, paternal semen, and maternal uterine tissue. Long-read libraries of the fetus were prepared for sequencing on the Sequel platform as suggested by the manufacturer (Pacific Biosciences, Menlo Park, CA). Short-read libraries of the sire and dam were prepared for sequencing on the NextSeq500 platform as recommended by the manufacturer (Illumina, Inc. San Diego, CA). Approximately 60x short-read coverage of the dam and sire were produced, and 134x long-read coverage for the fetus. Parent-specific kmers were identified, long reads sorted into bins by parental origin, and independent haploid assemblies constructed using triocanu (implemented in Canu v1.7).

Sex chromosomes assembly and validation

Haplotype resolved paternal and maternal contigs constructed using Canu [23] were scaffolded independently using Hi-C reads [49] and an optical map (Bionano tools v1.3.0), which were then consolidated

into chromosome specific groups of scaffolds and the sex chromosome scaffolds selected for the male and female assembly. The high density of repetitive elements on both the X and Y chromosomes made assembly difficult, breaking sequence contiguity even with long sequence reads, so additional markers were used to validate the order and orientation of scaffolds. For the X chromosome, the USDA-MARC Bovine linkage Map [35], and two RH maps, the BovGen RH map [37] and SUNbRH7000-rad map [36] were used to place, order and orientate scaffolds. For the Y chromosome, known genes in cattle [16, 22], pig (Sscrofa11.1) and human (GRCh38.p12) assemblies were used to identify Y-specific scaffolds. Cattle Y chromosome RH map markers [50] were used to guide ordering and orientation of the Y-specific scaffolds. To exclude scaffolds incorrectly identified as Y sequence, 26 scaffolds which were shorter than 50 kb and contained fewer than three known Y genes, were manually inspected. These shorter scaffolds were partitioned into 50-kb bins and aligned with the CHORI-240 Bovine BAC library Y specific clones (see URLs) using BLASTN. Six sequences with less 90% alignment were removed. Further details on X and Y chromosomes scaffolds identification and orientation is given in Additional file 1: Note 1 and 2, respectively.

Conflict resolution

The RH X chromosome marker order and orientation was generally in agreement with the X chromosome assembly. However, 15 out of 84 markers from the BovGen RH map and 18 out of 93 markers from the SUNbRH7000-rad map suggested a possible assembly error. Comparison of the RH maps showed that 5 markers from BovGen RH map and 4 markers from SUNbRH7000-rad map had a consistent arrangement that differed from the Brahman X assembly. Both RH maps showed an inversion of a 1,286,607 bp contig in a scaffold that is made up of 4 contigs. With the exception of this scaffold, other RH marker inconsistencies would have required contigs to be broken and rearranged. The Brahman X and ARS-UCD1.2 X agreed in all these regions so the order and orientation of scaffolds that agreed best with Hi-C and optical map was retained.

Outside of the PAR, the Y chromosome assembly is made up of small scaffolds with average length of 98,509 bp. The ordering of scaffolds was completely dependent on the 52 RH markers and no contig breaks were required.

X and Y chromosome alignments

In order to compare intra- and inter-species sex chromosomes differences, pairwise alignments were carried out for cattle (Brahman, Angus, Hereford, Nelore), water buffalo, goat, sheep, human, pig, dog and horse using the aligner Lastz v1.04 [51] with default settings.

Brahman X is reverse complement to Hereford X. Prior to alignments, the Hereford X and Nelore X are reverse orientated for comparison. Repeats in the X and Y chromosomes were masked by Repeatmasker v4–0-7 using cow RepBase23.08 [52]. The number of gaps was calculated using a custom python script. The number of inversions were calculated by Smash with a 1 Mb block size [53]. Further details on the alignment parameters is given in Additional file 1: Note 3.

X and Y chromosome gene annotations

All X chromosome genes from ARS-UCD1.2 were remapped to the Brahman X chromosome using the Exonerate v2.4 software [54]. For the PAR, detailed manual annotation was carried out. A total of 42 protein-coding PAR genes were used to annotate the X chromosome. These genes came from the following sources: 1) 15 human PAR1 genes [8]; 2) 22 candidate cattle PAR genes [11]; 3) horse [55], dog [56] and pig [57] PAR genes. A cattle PAR gene, arylsulfatase E (*ARSE*), was reported in a FISH mapping study [11] but the sequence is not available from NCBI. However, three other arylsulfatases paralogs on the X chromosome were available and were used to search for other similar members on the PAR.

The current cattle reference genome assembly ARS-UCD1.2 does not have a Y chromosome, therefore 18 Y-specific genes from the X-degenerate regions were selected from a study of Y chromosome transcribed sequences [16], in addition, MSY genes from pig and human Y chromosomes were used for the Angus Y chromosome annotation. Further details on gene annotation are given in Additional file 1: Note 4 and Additional file 1: Table S1.

Conservation of synteny in mammalian PARs

Each of the X and Y chromosomes PARs were contained in single scaffold. To estimate the boundary of PAR, X and Y scaffolds containing PARs were partitioned into 1 Mb windows and aligned against each other using BLASTN. From the alignments, the first bin from the proximal end that had coverage below 80% and percentage identity below 80% was identified as the pseudoautosomal boundary. The gene order of PAR genes between closely related species was analysed using Bioconductor Gviz plots [58].

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-019-6364-z>.

Additional file 1: Figure S1. Alignment of the Brahman X with other mammalian X chromosomes. **Figure S2.** Alignment of the Angus cattle Y against other mammalian Y chromosomes. **Figure S3.** Multi-copy genes in Angus Y ampliconic region **Table S1.** Summary of X and Y chromosome protein-coding genes. **Table S2.** Copy numbers of the protein-coding genes in the bovine MSY. **Table S3.** List of PAR genes missing in

the current Hereford, sheep, goat and pig assemblies.
Table S4. Copy numbers of OBP and BDA20 genes in each species.

Acknowledgements

We thank Alex Hastie for providing optical map based scaffolds.

Authors' contributions

JLW and TPLS conceived and managed the project. SH designed breeding experiments, provided dam mtDNA typing, tissue and DNA samples and clarified the single PAR in cattle using BAC information. TPLS created, sequenced, and curated sequencing data. SK, AR and AMP assembled the contigs and provided guidance on scaffolding. JG and SK provided the Hi-C scaffolds. DMB and BDR compared scaffolding using various programs. WYL consolidated all data on assemblies to produce final chromosome-level haplotype-resolved genomes. RL and BDR validated the sex chromosome assemblies using various bovine markers. RL annotated the sex chromosomes and compared them to other mammals. RL, WYL, RT, SH and JLW drafted the manuscript and all authors read, edited and approved the final manuscript.

Funding

Sample preparation and funding for JLW, SH, W-YL and RT was provided by the JS Davies Bequest to the University of Adelaide, and data production was funded from Project Number 3040–32000-034–00D of the United States Department of Agriculture Agricultural Research Service. A.R., S.K., and A.M.P. were supported by the Intramural Research Program of the National Human Genome Research Institute, US National Institutes of Health. A.R. was also supported by the Korean Visiting Scientist Training Award (KVSTA) through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (HI17C2098). Funders had no role in study design and interpretation of results.

Availability of data and materials

Sequence data is available from the BioProject Accession PRJNA432857. Analysis of Angus and Brahman genome assemblies, <https://github.com/loydlow/BrahmanAngusAssemblyScripts>; the CHORI-240 Bovine BAC Library is available from <https://bacpacresources.org/filtersAvail.php>; ArrowGrid, <https://github.com/skoren/ArrowGrid>; RepeatMasker, <http://www.repeatmasker.org>.

Ethics approval

All animal work was approved by the University of Adelaide Animal Ethics Committee, approval No. S-094-2005.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia. ²Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA. ³Center for Bioinformatics and Computational Biology, Lab 3104A, Biomolecular Science Building, University of Maryland, College Park, MD, USA. ⁴Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, MD, USA. ⁵Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, WI, USA. ⁶US Meat Animal Research Center, ARS USDA, Clay Centre, NE, USA.

Received: 26 July 2019 Accepted: 2 December 2019

Published online: 19 December 2019

References

- Ohno S. Evolution of sex chromosomes in mammals. *Annu Rev Genet.* 1969;3(1):495–524.
- Bull JJ. Evolution of sex determining mechanisms: the Benjamin/Cummings publishing company, Inc; 1983.
- Graves JA. Sex chromosome specialization and degeneration in mammals. *Cell.* 2006;124(5):901–14.

4. Charlesworth B. The evolution of sex chromosomes. *Science*. 1991; 251(4997):1030–3.
5. Graves JA. Evolution of the mammalian Y chromosome and sex-determining genes. *J Exp Zool*. 1998;281(5):472–81.
6. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*. 2003;423(6942):825–37.
7. NaG E-M. J.A.M: X and Y Chromosomes: Homologous Regions. Chichester: eLS John Wiley & Sons Ltd; 2008.
8. Helena Mangs A, Morris BJ. The human Pseudoautosomal region (PAR): origin, Function and Future. *Curr Genomics*. 2007;8(2):129–36.
9. Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*. 2005;437(7055):100–3.
10. Van Laere AS, Coppieters W, Georges M. Characterization of the bovine pseudoautosomal boundary: documenting the evolutionary history of mammalian sex chromosomes. *Genome Res*. 2008;18(12):1884–95.
11. Das PJ, Chowdhary BP, Raudsepp T. Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenet Genome Res*. 2009;126(1–2):139–47.
12. Johnson T, Keehan M, Harland C, Lopdell T, Spelman RJ, Davis SR, Rosen BD, Smith TPL, Coudrey C. Short communication: identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. *J Dairy Sci*. 2019;102(4):3254–8.
13. Gabriel-Robez O, Rumpel Y, Ratomponirina C, Petit C, Levilliers J, Croquette MF, Couturier J. Deletion of the pseudoautosomal region and lack of sex-chromosome pairing at pachytene in two infertile men carrying an X,Y translocation. *Cytogenet Cell Genet*. 1990;54(1–2):38–42.
14. Mohandas TK, Speed RM, Passage MB, Yen PH, Chandley AC, Shapiro LJ. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *Am J Hum Genet*. 1992;51(3):526–33.
15. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*. 2010;463(7280):536–9.
16. Chang TC, Yang Y, Retzel EF, Liu WS. Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc Natl Acad Sci U S A*. 2013;110(30):12373–8.
17. Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature*. 2012;483(7387):82–6.
18. Soh YQ, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell*. 2014;159(4):800–13.
19. Skinner BM, Sargent CA, Churcher C, Hunt T, Herrero J, Loveland JE, Dunn M, Louzada S, Fu B, Chow W, et al. The pig X and Y chromosomes: structure, sequence, and evolution. *Genome Res*. 2016;26(1):130–9.
20. Janecka JE, Davis BW, Ghosh S, Paria N, Das PJ, Orlando L, Schubert M, Nielsen MK, Stout TAE, Brashear W, et al. Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nat Commun*. 2018;9(1):2945.
21. Liu Y, Qin X, Song XZ, Jiang H, Shen Y, Durbin KJ, Lien S, Kent MP, Sodeland M, Ren Y, et al. Bos taurus genome assembly. *BMC Genomics*. 2009;10:180.
22. Canavez FC, Luche DD, Stothard P, Leite KR, Sousa-Canavez JM, Plastow G, Meidanis J, Souza MA, Feijao P, Moore SS, et al. Genome sequence and assembly of *Bos indicus*. *J Hered*. 2012;103(3):342–8.
23. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, TPL S, Phillippy AM. De novo assembly of haplotypesolved genomes with trio binning. *Nature Biotechnol*. 2018;36(12):1174.
24. Hiendleder S, Lewalski H, Janke A. Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. *Cytogenet Genome Res*. 2008;120(1–2):150–6.
25. Consortium BG, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*. 2009;324(5926):522–8.
26. Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. Evidence for two independent domestications of cattle. *Proc Natl Acad Sci U S A*. 1994; 91(7):2757–61.
27. LWPdL FA. Mapping of the bovine Y chromosome. *Electron J Biol*. 2007;3(1):8.
28. Kuderna LFK, Lizano E, Julia E, Gomez-Garrido J, Serres-Armero A, Kuhlwlilm M, Alandes RA, Alvarez-Estape M, Juan D, Simon H, et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nat Commun*. 2019;10(1):4.
29. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun*. 2019;10(1):260.
30. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet*. 2017;49(4):643–50.
31. Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, et al. The yak genome and adaptation to life at high altitude. *Nat Genet*. 2012;44:946.
32. Bana NA, Nyiri A, Nagy J, Frank K, Nagy T, Steger V, Schiller M, Lakatos P, Sugar L, Horn P, et al. The red deer *Cervus elaphus* genome CerElal.0: sequencing, annotating, genes, and chromosomes. *Mol Gen Genomics*. 2018;293(3):665–84.
33. Ge RL, Cai Q, Shen YY, San A, Ma L, Zhang Y, Yi X, Chen Y, Yang L, Huang Y, et al. Draft genome sequence of the Tibetan antelope. *Nat Commun*. 2013;4:1858.
34. International Sheep Genomics C, Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, JC ME, Hutton Oddy V, Raadsma HW, et al. The sheep genome reference sequence: a work in progress. *Anim Genet*. 2010;41(5):449–53.
35. Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M, Mizoguchi Y, Hirano T, Itoh T, Watanabe T, Reed KM, et al. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res*. 2004; 14(10A):1987–98.
36. Itoh T, Watanabe T, Ihara N, Mariani P, Beattie CW, Sugimoto Y, Takasuga A. A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics*. 2005;85(4):413–24.
37. Jann OC, Aerts J, Jones M, Hastings N, Law A, McKay S, Marques E, Prasad A, Yu J, Moore SS, et al. A second generation radiation hybrid map to aid the assembly of the bovine genome sequence. *BMC Genomics*. 2006;7:283.
38. Iannuzzi L, King WA, Di Berardino D. Chromosome evolution in domestic bovids as revealed by chromosome banding and FISH-mapping techniques. *Cytogenet Genome Res*. 2009;126(1–2):49–62.
39. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, Timetrees, and divergence times. *Mol Biol Evol*. 2017;34(7):1812–9.
40. Arnason U, Janke A. Mitogenomic analyses of eutherian relationships. *Cytogenet Genome Res*. 2002;96(1–4):20–32.
41. Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, Takahata N. The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proc Natl Acad Sci U S A*. 2003;100(9):5258–63.
42. Logdberg L, Wester L. Immunocalins: a lipocalin subfamily that modulates immune and inflammatory responses. *Biochim Biophys Acta*. 2000;1482(1–2):284–97.
43. Wang YH, Reverter A, Kemp D, McWilliam SM, Ingham A, Davis CA, Moore RJ, Lehnert SA. Gene expression profiling of Hereford shorthorn cattle following challenge with *Boophilus microplus* tick larvae. *Aust J Exp Agric*. 2007;47(12):1397–407.
44. Rice WR. Evolution of the Y sex chromosome in animals. *Bioscience*. 1996; 46(5):331–43.
45. Rozen S, Warren WC, Weinstock G, O'Brien S. J GRWRKPD: sequencing and annotating new mammalian Y chromosomes a white paper proposal; 2006.
46. Liu WS, Zhao Y, Lu C, Ning G, Ma Y, Diaz F, O'Connor M. A novel testis-specific protein, PRAMEY, is involved in spermatogenesis in cattle. *Reproduction*. 2017;153(6):847–63.
47. Yang Y, Chang TC, Yasue H, Bharti AK, Retzel EF, Liu WS. ZNF280BY and ZNF280AY: autosome derived Y-chromosome gene families in Bovidae. *BMC Genomics*. 2011;12:13.
48. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*. 2014;508(7497):494–9.
49. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *bioRxiv* 2019. p. 261149.

50. Liu WS, Mariani P, Beattie CW, Alexander LJ, Ponce De Leon FA. A radiation hybrid map for the bovine Y chromosome. *Mamm Genome*. 2002;13(6): 320–6.
51. Harris RS. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University; 2007.
52. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11.
53. Pratas D, Silva RM, Pinho AJ, Ferreira PJ. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. *Sci Rep*. 2015;5:10203.
54. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31.
55. Moore SS, Byrne K, Johnson SE, Kata S, Womack JE. Physical mapping of CSF2RA, ANT3 and STS on the pseudoautosomal region of bovine chromosome X. *Anim Genet*. 2001;32(2):102–4.
56. Young AC, Kirkness EF, Breen M. Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: the canine PAR and PAB. *Chromosom Res*. 2008;16(8):1193–202.
57. Skinner BM, Lachani K, Sargent CA, Affara NA. Regions of XY homology in the pig X chromosome and the boundary of the pseudoautosomal region. *BMC Genet*. 2013;14:3.
58. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. In: Mathé E, Davis S, editors. *Statistical Genomics: Methods and Protocols*. New York: Springer New York; 2016. p. 335–51.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



2.1 Supplementary Notes

2.1.1 X chromosome scaffolds identification and orientation

X chromosome RH and linkage map markers (Ihara et al., 2004; Itoh et al., 2005; Jann et al., 2006) were aligned to the maternal (Brahman) contigs using `blastn` (Camacho et al., 2009) with the following parameters: (`-max_hsps 3 -perc_identity 90 -qcov_hsp_perc 80`). Contigs were ordered based on the linkage and RH map marker order and then used to assess the concordance with scaffolds based on Hi-C and optical maps. Three Hi-C scaffolds were broken and re-joined based on the RH map order. There was no conflict between optical map and RH map. The Brahman X chromosome was scaffolded by optical mapping (17 scaffolds consisting of 107 contigs) and by Hi-C (14 scaffolds consisting of 57 contigs). The additional optical mapping contigs had a total length of 10 Mb. Hi-C scaffolds were longer, on average, than optical map based scaffolds. As the optical map based scaffolds were in closer agreement with the linkage and RH data, their scaffolds were joined together based on Hi-C scaffolds. This was then polished, gap filled (English et al., 2012) and the consensus sequence rebuilt with ArrowGrid (<https://github.com/skoren/ArrowGrid>). The final length of the X chromosome was 146,092,946 Mb.

2.1.2 Y chromosome scaffolds identification and orientation

The Y chromosome contigs were initially identified by the presence of genes reported on the Y chromosomes on the bovine genome assemblies (Bos indicus 1.0, Btau 5.0.1) and a previous cattle Y chromosome gene expression study (Chang et al., 2013). We then aligned Y chromosome SNP probes from bovine HD 50k chip and 62 Y-linked RH map markers (Liu et al., 2002; Stafuzza et al., 2009) to paternal contigs and confirmed that 91 contigs originated from cattle Y chromosome. Six contigs with small number of Y chromosome SNP probes were not included as Y sequence because they were in conflict with optical map scaffolding, which placed them on large 4 autosomal scaffolds. In rest of 73 contigs, six contigs were removed as they have <80% sequencing identify in alignment with CHORI-240 Bovine BAC library Y. The Hi-C and optical map scaffolds were ordered and orientated based on RH map markers to produce scaffolds that were in best agreement.

The final length of the Y chromosome was 15,658,480 Mb. This was then subjected to gap filling (English et al., 2012) and consensus sequence rebuilding with ArrowGrid.

2.1.3 Comparison of X and Y chromosomes in mammals

To facilitate comparison of sex chromosomes across mammalian species representative reference genomes were downloaded from NCBI, these were X (cattle, water buffalo, sheep, human, pig, dog horse) and Y chromosome (cattle, human, pig) were. For the goat X chromosome, we manually joined two X chromosome unplaced scaffolds in the goat genome (NW_017189516.1 and NW_017189517.1) and aligned them to our Brahman X. Horse Y chromosome sequences were obtained from Horse eMSYv3.1 assembly (GenBank MH341179) (Janecka et al., 2018). Prior to alignments, repeats in the Brahman X and Angus Y chromosomes were masked by Repeatmasker v4-0-7 using cow RepBase23.08 (Bao et al., 2015). Repeat-masked sex chromosomes from other species were downloaded from the NCBI. Pairwise alignments were generated using the aligner Lastz v1.04 (Harris, 2007) with the following parameters.

(i) For intra species:

```
: --notransition --step=20 --nogapped --format=maf --
  ambiguous=iupac\
```

(ii) For inter species:

```
--notransition --step=50 --nogapped --format=maf --
  ambiguous=iupac\
```

2.1.4 Gene annotation of sex chromosomes

To annotate the X chromosome, we downloaded mRNA sequences from the cattle assemblies in NCBI (ARS-UCD 1.2 and *Bos indicus* 1.0) and lifted these over to the Brahman X using Exonerate v2.4.0 (Slater and Birney, 2005) with a cut off of 88% (parameters:

```
--model est2genome --querytype dna--targettype dna--showvulgar no --
  showalignment no --showtargetgff yes --showcigar no --percent 88
```

). A total of 983 genes were annotated on the Brahman X. To annotate the Y chromosome, we downloaded the mRNA from cattle Y chromosome (Btau 5.0.1), previous cattle Y chromosome sequence (Chang et al., 2013) and other mammalian orthologous sequences (human, pig and horse) and aligned these against the Angus Y assembly. We lifted cattle Y chromosome sequences over using same parameters as we used for Brahman X. In addition to cattle Y genes, other homologous Y genes from human, pig, and horse were used as input to search for Y genes in Angus using 75% sequence identity as cut off. In total, 51 unique genes were annotated in Angus Y including genes on PAR.

2.2 Supplementary Figures

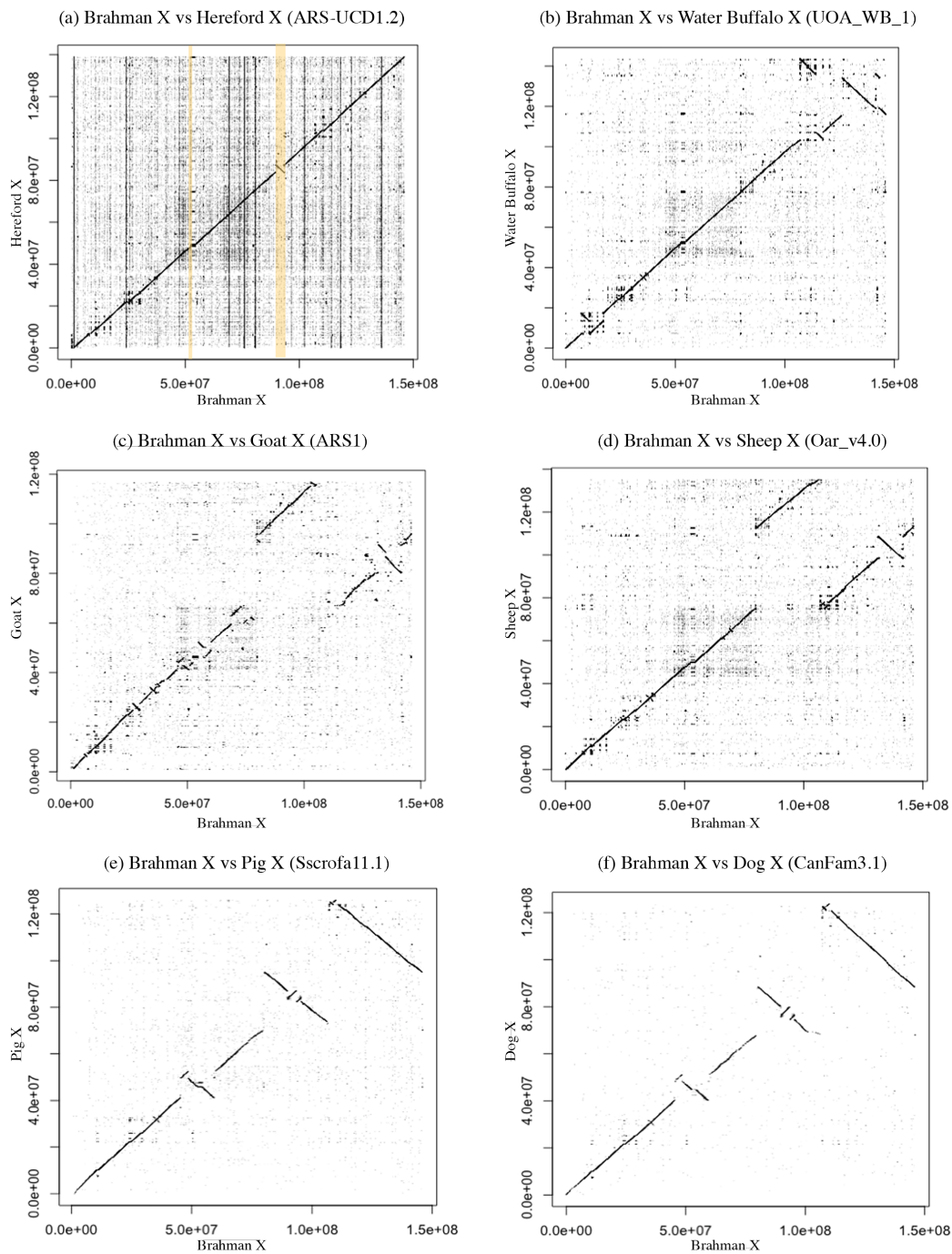


Figure 2.1: Alignment of the Brahman X with other mammalian X chromosomes.

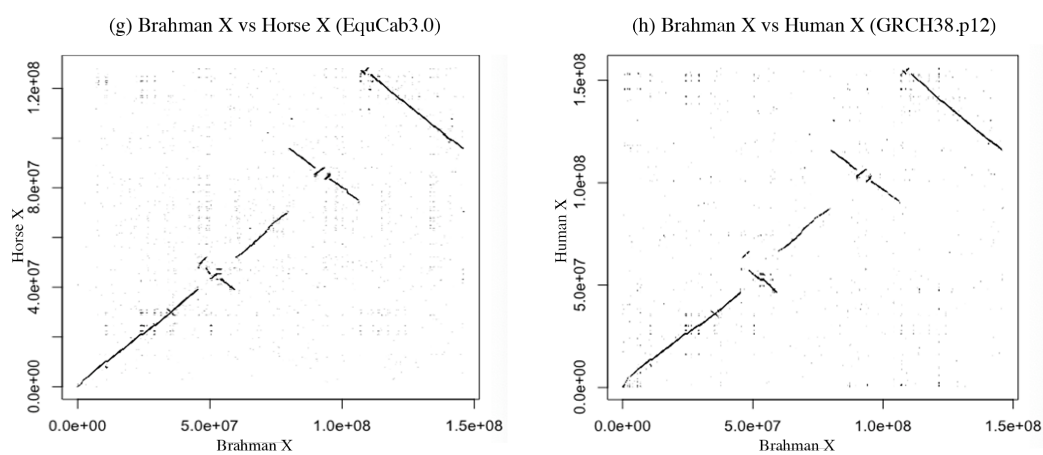


Figure 2.1: (continued) Alignment of the Brahman X with other mammalian X chromosomes.

The Brahman X chromosome is displayed on the x-axes and the X chromosomes from other species on the y-axes. The panels show alignments with a) Cattle (Hereford) X from ARS-UCD1.2 assembly. Two major inversion blocks (>1 Mb) are highlighted in yellow. b) Water Buffalo X from UOA_WB_1 assembly. c) Goat X from ARS1 assembly. d) Sheep X from Oar_v4.0 assembly. e) Pig X from Sscrofa11.1 assembly. f) Dog X from CanFarm3.1 assembly. g) Horse X from EquCab3.0 assembly. h) Human X from GRCH38.p12 assembly.

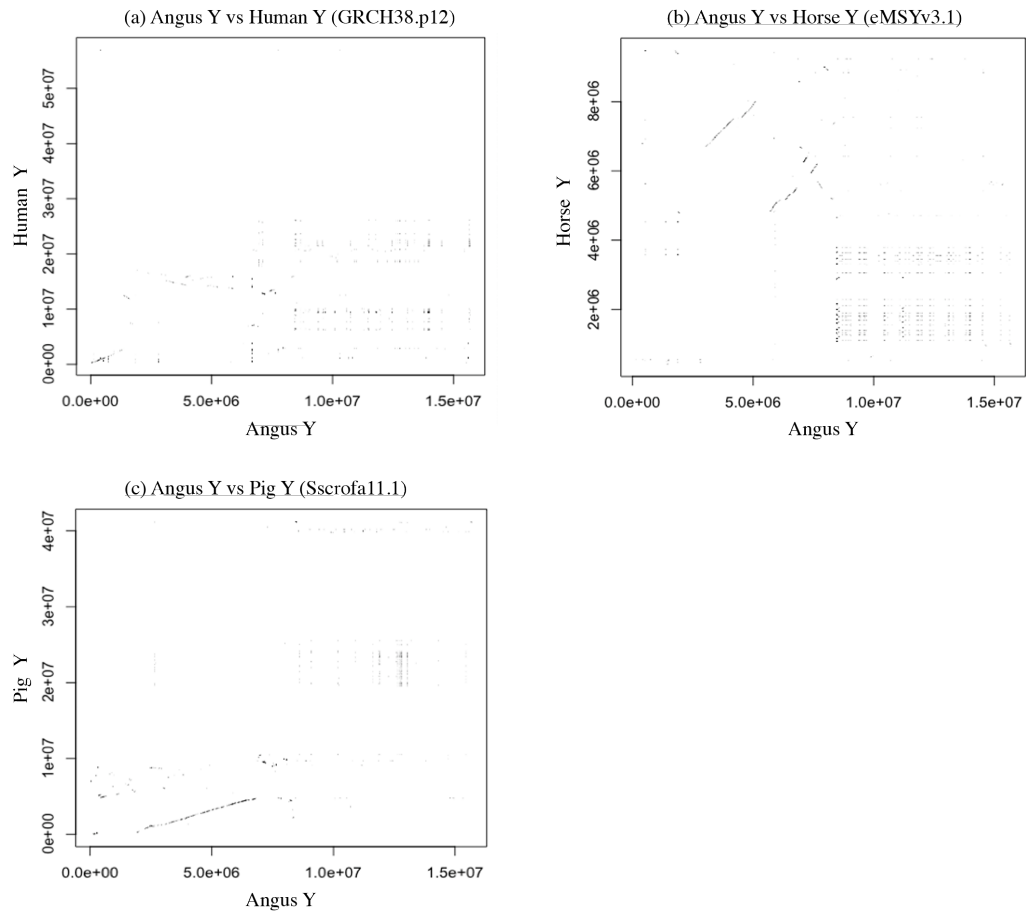


Figure 2.2: Alignment of the Angus cattle Y against other mammalian Y chromosomes.

The Angus Y chromosome is on the x-axis and the Y chromosomes from other species on the y-axis. Panels are a) Human Y from GRCH38.p12 assembly. b) Horse Y from eMSYv3.1 assembly. c) Pig Y from Sscrofa11.1 assembly.

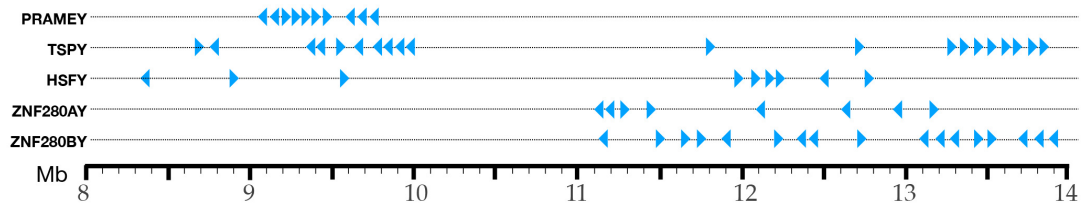


Figure 2.3: Multi-copy genes in Angus Y ampliconic region.

The ampliconic region are from 8.07Mb to 10.01Mb and 10.81Mb to 14.05Mb. The triangular dot plot showing the location of gene copies in Angus Y ampliconic region.

2.3 Supplementary Tables

Table 2.1: Summary of X and Y chromosome protein-coding genes

X Genes	Total	Single copy	Multi-copy
X-specific	936	936	
PAR	31	29	2
X-Y paired outside PAR	16	16	-
Total genes	983	981	2
Total transcripts	988	981	7
Y Genes	Total	Single copy	Multi-copy
Y-specific	4	0	4
PAR	31	29	2
X-Y paired outside PAR	15	14	1
Total genes	51	43	6
Total transcripts	153	43	113

A total of 983 genes were annotated on the X chromosome by lift over from three cattle assemblies (ARS-UCD1.2, Bos indicus 1.0 and Btau 5.0.1). These included 936 X-specific single/multiple copy genes, 31 PAR genes and 16 X-Y paired genes. Lift over of Y chromosome annotated genes from previous cattle Y chromosome sequences (Chang et al., 2013) identified a total of 51 genes including 31 PAR genes, 15 X-Y paired genes and four Y-specific genes without X-homologues. Total number genes on Y including PAR genes are 153. These genes were used as input for Exonerate v2.4.0 to search the Brahman X and Angus Y chromosomes.

Table 2.2: Summary of X and Y chromosome protein-coding genes

Y Genes (X-d region)	Total
EIF1AY	1
AMELY	1
OFD1Y	1
USP9Y	1
ZRSR2Y	1
UTY	1
DDX3Y	1
SHROOM2Y	1
ZFY	1
EIF2S3Y	1
UBE1Y	1
TXLNGY	1
SRY	1
RBMX	1
Y Genes (ampliconic region)	Total
PRAMEY	10
HSFY	9
TSPY	20
ZNF280AY	8
ZNF280BY	17

This is the summary of copy number of known MSY genes. there are 15 genes in X-d regions and 5 known multi-copy genes in ampliconic region.

Table 2.3: Summary of X and Y chromosome protein-coding genes

PAR genes	Hereford		Sheep		Goat		Pig	
	Chr	Accession version	Chr	Accession version	Chr	Accession version	Chr	Accession version
PLCXD1	Chr1	NM_001105044.1	ChrUn	XM_004022871.2	ChrUn	XM_018044435.1	-	-
GTPBP6	Chr1	XM_024995411.1	ChrUn	XM_012107966.1	ChrUn	XM_018044432.1	-	-
PPP2R3B	Chr1	XM_024995403.1	ChrUn	XM_012107967.1	ChrUn	XM_018044433.1	ChrUn	XM_021081364.1
SHOX	Chr2	NM_001191546.2	-	-	ChrUn	XM_018045110.1	ChrUn	XM_021081396.1
CRLF2	Chr29	XM_024987475.1	✓	✓	ChrUn	XM_005701362.3	ChrUn	XM_021081385.1
CSF2RA	Chr3	XM_024990295.1	✓	✓	ChrUn	XM_005701352.3	ChrUn	XM_021081383.1
IL3RA	Chr3	XM_024990300.1	✓	✓	ChrUn	XM_018044557.1	ChrUn	XR_002341032.1
SLC25A6	Chr26	NM_174660.2	✓	✓	ChrUn	XM_018044555.1	ChrUn	NM_214418.2
ASMTL	Chr26	NM_001035058.1	✓	✓	ChrUn	XM_018044443.1	ChrUn	XM_021081390.1
P2RY8	Chr3	XM_005228542.3	✓	✓	ChrUn	XM_013976840.2	ChrUn	XM_021081386.1
AKAP17A	Chr3	XM_024990316.1	✓	✓	ChrUn	XM_018044444.1	ChrUn	XM_021081363.1
ASMT	Chr1	NM_177493.2	✓	✓	ChrUn	NM_001285598.1	ChrUn	XM_021081386.1
DHRX	✓	✓	✓	✓	ChrUn	XM_018044442.1	ChrY	XM_021080887.1
ZBED1	✓	✓	✓	✓	ChrUn	XM_018044439.1	ChrUn	XM_021082454.1
CD99	✓	✓	✓	✓	✓	✓	ChrUn	XR_002340897.1
XG	✓	✓	✓	✓	✓	✓	✓	✓
GYG2	✓	✓	✓	✓	✓	✓	✓	✓
ARSD	✓	✓	✓	✓	✓	✓	✓	✓
ARSE	✓	✓	✓	✓	✓	✓	✓	✓
ARSH	✓	✓	✓	✓	✓	✓	✓	✓

The PAR genes missing from the cattle (Hereford) assembly were found on five autosomes (chromosome 1,2,3,26, 29). Missing PAR genes in the in sheep, goat and pig assemblies were found either in unplaced scaffolds or as follows: sheep SHOX and pig GTPBP6, PPP2R3B and ARSH were not found in their current X chromosome assembly. IL3RA and CD99 are among the non-coding RNA in the current pig assembly.

Table 2.4: Copy numbers of OBP and BDA20 genes in each species

	Brahman	Hereford	Water buffalo	Sheep	Goat	Pig	Dog	Horse
OBP	ENSBIXG0000 5001147	XM_024988 434.1	XM_025276 408.1	XM_027963 189.1	XM_018043 719.1	NM_213796.1	XM_025440 685.1	XM_014728 749.1
	ENSBIXG0000 5011676	XM_005228 543.4	XM_025276 466.1	XM_027963 188.1	XM_018044 049.1	XM_021080 597.1	XM_005640 968.2	XM_014728 750.1
	ENSBIXG0000 5001130	XM_002700 469.6	XM_025276 502.1	-	-	-	XM_025435 144.1	-
	-	XM_024988 435.1	XM_025276 828.1	-	-	-	-	-
	-	XM_010800 829.3	-	-	-	-	-	-
BDA20	ENSBIXG0000 5011448	XM_024988 433.1	XM_006074 786.2	XM_012106 177.2	XM_018043 718.1	-	-	-
	ENSBIXG0000 5011472	XM_010822 282.3	XM_025276 762.1	XM_006074 786.2	XM_018044 027.1	-	-	-
	ENSBIXG0000 5011425	XM_010800 830.3	XM_025276 829.1	-	XM_005701 239.2	-	-	-
	ENSBIXG0000 5011448	-	-	-	XM_018044 047.1	-	-	-

There are varies copy number of OBP and BDA20 among species. OBP found in all species we compared in this paper except human. BDA20 only found in ruminant species. 2 copies in sheep, three copies in Goat and 4 copies in both brahman and Hereford.

References

- Bao, W., Kojima, K.K., and Kohany, O., 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob dna* [Online], 6, p.11.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L., 2009. Blast+: architecture and applications. *Bmc bioinformatics* [Online], 10(1), p.421.
- Chang, T.C., Yang, Y., Retzel, E.F., and Liu, W.S., 2013. Male-specific region of the bovine y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc natl acad sci u s a* [Online], 110(30), pp.12373–8.
- English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A., 2012. Mind the gap: upgrading genomes with pacific biosciences rs long-read sequencing technology. *Plos one* [Online], 7(11), e47768.
- Harris, R.S., 2007. *Improved pairwise alignment of genomic dna*. Thesis.
- Ihara, N., Takasuga, A., Mizoshita, K., Takeda, H., Sugimoto, M., Mizoguchi, Y., Hirano, T., Itoh, T., Watanabe, T., Reed, K.M., Snelling, W.M., Kappes, S.M., Beattie, C.W., Bennett, G.L., and Sugimoto, Y., 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome res* [Online], 14(10A), pp.1987–98.
- Itoh, T., Watanabe, T., Ihara, N., Mariani, P., Beattie, C.W., Sugimoto, Y., and Takasuga, A., 2005. A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics* [Online], 85(4), pp.413–24.
- Janecka, J.E., Davis, B.W., Ghosh, S., Paria, N., Das, P.J., Orlando, L., Schubert, M., Nielsen, M.K., Stout, T.A.E., Brashear, W., Li, G., Johnson, C.D., Metz, R.P., Zadjali, A.M.A., Love, C.C., Varner, D.D., Bellott, D.W., Murphy, W.J., Chowdhary, B.P., and Raudsepp, T., 2018. Horse y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nat commun* [Online], 9(1), p.2945.

- Jann, O.C., Aerts, J., Jones, M., Hastings, N., Law, A., McKay, S., Marques, E., Prasad, A., Yu, J., Moore, S.S., Floriot, S., Mahe, M.F., Eggen, A., Silveri, L., Negrini, R., Milanesi, E., Ajmone-Marsan, P., Valentini, A., Marchitelli, C., Savarese, M.C., Janitz, M., Herwig, R., Hennig, S., Gorni, C., Connor, E.E., Sonstegard, T.S., Smith, T., Drogemuller, C., and Williams, J.L., 2006. A second generation radiation hybrid map to aid the assembly of the bovine genome sequence. *Bmc genomics* [Online], 7, p.283.
- Liu, W.S., Mariani, P., Beattie, C.W., Alexander, L.J., and Ponce De Leon, F.A., 2002. A radiation hybrid map for the bovine y chromosome. *Mamm genome* [Online], 13(6), pp.320–6.
- Slater, G.S. and Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *Bmc bioinformatics* [Online], 6, p.31.
- Stafuzza, N.B., Abbassi, H., Grant, J.R., Rodrigues-Filho, E.A., Ianella, P., Kadri, S.M., Amarante, M.V., Stohard, P., Womack, J.E., Leon, F.A. de, and Amaral, M.E., 2009. Comparative rh maps of the river buffalo and bovine y chromosomes. *Cytogenet genome res* [Online], 126(1-2), pp.132–8.

3 X-Y chromosome gametologues explain sex differences in fetal organ weights

Ruijie Liu¹, Rick Tearle¹, Wai Yee Low¹, Tong Chen¹, Dana Thomsen^{1,2}, Consuelo Amor S. Estrella^{1,2,4}, Ruidong Xiang^{1,2,3}, David R. Rutley¹, Timothy P.L. Smith⁵, John L. Williams¹, Stefan Hiendleder^{1,2}

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia

²Robinson Research Institute, The University of Adelaide, Adelaide, Australia

³Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, Australia

⁴University of the Philippines, Laguna, Philippines

⁵USMARC, USDA-ARS-US Meat Animal Research Center, Clay Center, NE, USA

Intended for submission to Nature Genetics in Brief Communication format

3.1 Abstract

Gametologues are homologous non-recombining genes on the X and Y chromosomes proposed to contribute to phenotypic differences between the sexes. We show that a remarkably small set of gametologues distinguishes male and female transcriptomes across fetal tissues and placenta. The dosage of differentially expressed X-Y paired gametologues in females and males is frequently unbalanced and explains 18% - 96% of the phenotypic variance in organ weights attributed to the sex effect.

3.2 Main

Phenotypic differences between females and males arise early in embryonic development, before the onset of hormone production in the developing gonads (Arnold, 2017). In mammals, early sex-differences result from the inherent genetic differences between the female XX and male XY chromosome complements (Snell and Turner, 2018; Arnold, 2019). Evolutionary studies have previously identified conserved X-Y chromosome paired gametologues outside the pseudo-autosomal region (PAR), which escape X chromosome inactivation (XCI) in females and function as widely expressed regulators of gene expression, forming candidate genes for phenotypic differences between the sexes (Bellott et al., 2014).

Here we use a bovine mid-gestation (Day 153) fetal model as it enters accelerated growth and female-male phenotypic differentiation (Xiang et al., 2013; Xiang et al., 2014), in combination with high quality bovine sex chromosome (Liu et al., 2019) and autosome assemblies (Low et al., 2020), to demonstrate a clear separation of female and male transcriptomes in brain, liver, lung, skeletal muscle and fetal placenta (Supplementary Tables 3.1-3.2, Supplementary Figure 3.3). We identified 54 genes (28 X-chromosome and 10 Y-chromosome linked, 16 autosomal) that were differentially expressed (DE) between females and males in at least one tissue (Supplementary Table 3.2, Supplementary Figure 3.4). The comparatively low number of DE genes detected here is similar to results obtained for the mouse embryo (Lowe et al., 2015), but contrasts sharply with hundreds to thousands of sex-dependent DE genes in adult mammalian tissues (Yang et al., 2006; Seo

et al., 2016; Gershoni and Pietrokovski, 2017), including our comparison of fetal and adult bovine liver and muscle (Supplementary Figure 3.5); an increase that can be attributed to the effects of sex hormones (Snell and Turner, 2018; Arnold, 2019). Importantly, 16 of the

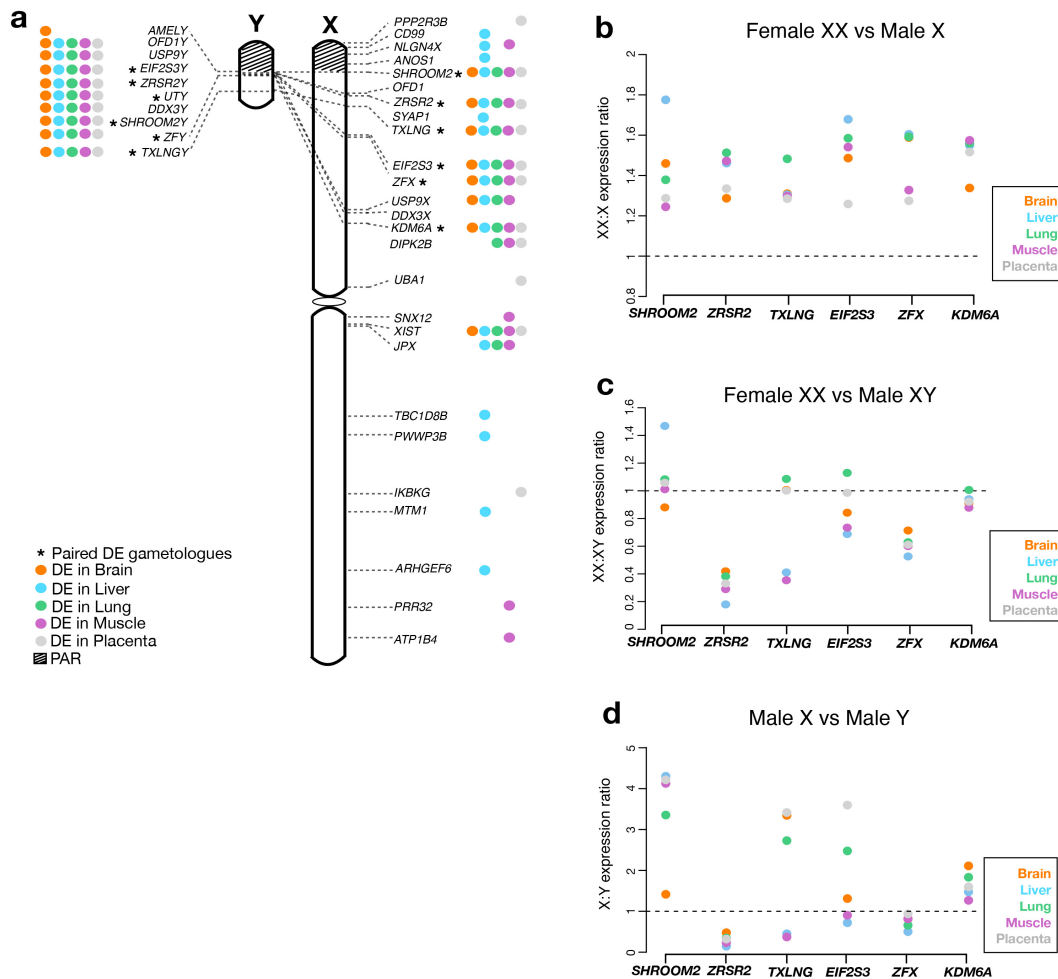


Figure 3.1: Differentially expressed genes on the fetal sex chromosomes discriminating females and males at midgestation (Day153).

Differentially expressed genes on the fetal sex chromosomes discriminating females and males at midgestation (Day153). a) Ideograms of the X and Y chromosomes with 34 genes differentially expressed (DE) between females and males. Tissue specificity is indicated by coloured circles. Among a total of 16 sex-specific DE genes that were shared across transcriptomes of all 5 investigated tissues, all but XIST are gametologues. Asterisks mark gametologues where both the X and Y homologue are DE; these paired genes were further analysed in b-d. b) Expression ratios of DE X chromosome homologues in females and males are >1 and thus consistent with partial escape of one female X homologue from X chromosome inactivation (XCI). c) Expression ratios of female XX and combined male XY gametologues are <1 for the majority (22/30) of gametologue pairs, indicating a male bias due to incomplete escape of one homologue from XCI in females. d) Most expression ratios of X and Y homologues in males are unbalanced, and may thus explain some XX:XY gametologue expression ratios that are balanced or even female biased.

DE fetal genes were shared across all analysed tissues (Supplementary Figure 3.4), and all were located on the sex chromosomes (Figure 3.1a). This set includes the X-inactivating *XIST*, an essential component of dosage compensation that corrects the imbalance in the number of X chromosomes between females and males in fetal (Supplementary Figure 3.6) and adult (Sahakyan et al., 2018) tissues, and 15 gametologues with essential roles in chromatin modification, transcription, translation and protein stability (Bellott et al., 2014). Twelve of these gametologues comprise pairs where both the X and Y homologue are DE (*KDM6A* aka *UTX/UTY*, *ZRSR2/ZRSR2Y*, *EIF2S3/EIF2S3Y*, *ZFX/ZFY*, *TXLNG/TXLNGY*, *SHROOM2/SHROOM2Y*), while the remaining three (*DDX3Y*, *OFD1Y*, *USP9Y*) are DE only from the Y homologue (Figure 3.1a, Supplementary Table 3.2).

We focused further analyses on gametologue expression and its role in phenotypic sex differences on the 6 pairs with DE X and Y homologues, as XX:XY dosage ratio (see below) was inherently male biased when only the Y homologue is DE. Our initial comparison of the expression of X chromosome homologues in females and males revealed tissue-specific expression ratios of >1 but <2 (range ~ 1.2 - 1.8) that are consistent with incomplete escape of one of the two female X homologues from XCI (Figure 3.1b). We next compared the female XX and combined male XY gametologue dosage ratios and found that the majority of gametologues and tissues display a ratio of <1 (22/30), indicating a male bias (Figure 3.1c). We conclude that, like in adult human (Tukiainen et al., 2017), XCI effects prevent the full escape of homologues on the inactive X chromosome (X_i) of the bovine fetus. This is further supported by apparent XCI effects on PAR genes (Tukiainen et al., 2017) that causes male biased expression (Supplementary Figure 3.7). However, XCI effects on XX:XY gametologue dosage ratio are frequently modified by under- or overexpression of the Y-homologue relative to the male X-homologue (Figure 3.1d). This impacts the magnitude of male bias in fetal XX:XY dosage ratio and can, in a few instances, even shift the ratio to balanced or in favour of the female (Figure 3.1c).

The comparison of fetal and adult female-male DE genes in both liver and muscle revealed that apart from *XIST* only the 9 Y-linked gametologues (*DDX3Y*, *EIF2S3Y*, *KDMA6D*, *OFD1Y*, *SHROOM2Y*, *TXLNGY*, *ZFY*, *ZRSR2Y*) are shared across developmental stages

(Supplementary Figure 3.8). Detailed analyses of the expression of X and Y homologues revealed substantial ontogenetic changes with a general shift to more balanced XX:XY dosage ratios, particularly in muscle (Supplementary Figure 3.9). The involvement of DE X chromosome homologues in gametologue dosage of fetal but not adult tissues supports the view that important female-male differences in postnatal phenotype are programmed in the fetus (Gabory et al., 2013).

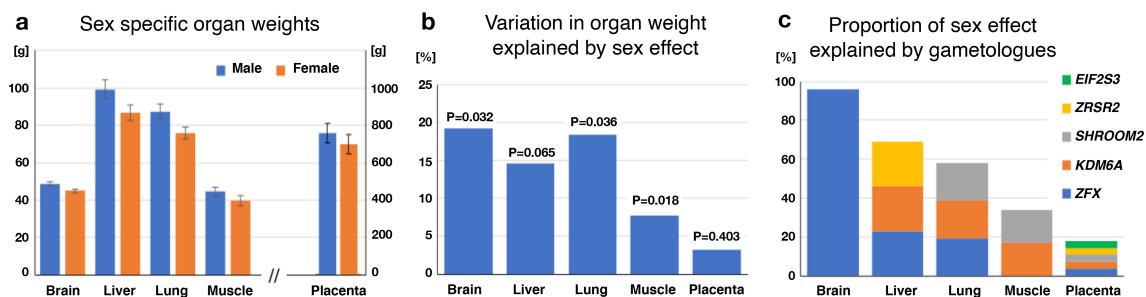


Figure 3.2: Female-male differences in gametologue expression explain sex effects on fetal organ weights at midgestation (Day153).

a) Mean \pm standard error of female and male organ weights as determined by ANOVA, with fetus numbers given in bars. b) Proportion of the phenotypic variance in organ weight explained by the sex effect in ANOVA. c) Proportion of the phenotypic variance due to the sex effect that is explained by differences in gametologue expression of females and males. Note: The actual relative proportions of the effects of each gametologue for explaining variation due to sex have not been determined.

The mid-gestation fetuses analysed here begin to display sex differences in phenotype, including heavier (ANOVA, $P < 0.05$) brain, lung and skeletal muscle weights of males (Figure 3.2a). We therefore developed an approach based on linear models to test whether differences in female and male gametologue dosage can explain sex-specific phenotypic variation in organ weights (Figure 3.2b). We found that gametologues account for 18% - 96% of the variation attributed to the sex effect in ANOVA (Figure 3.2c, Supplementary Tables 3.4-3.5). Notably, *ZFX/ZFY* dosage alone explains 96% of the sex variation in brain weight, while combinations of 2-5 gametologues explain up to 69% of the sex variation in other organ weights. The zinc-finger protein encoded by *ZFX/ZFY* is a transcriptional regulator with important functions in stem cell renewal, cell cycle regulation and growth control; it promotes tumour growth in various tissues, including brain (Huret et al., 2013). While the XX:XY dosage ratio for *ZFX/ZFY* shows a clear male bias, ratios for other

gametologues contributing to sex variation in organ weights are not always unbalanced (Figure 3.1c). An example is *KDM6A* aka *UTX/UTY*, which, like *ZFX/ZFY*, contributes to sex differences in 4 organs (Figure 3.2c), but whose XX:XY dosage in 3 affected organs is balanced or close to balanced (Supplementary Table 3.6). The X chromosome homologue *KDM6A* encodes a histone demethylase involved in transcriptional activation and regulation of growth and development (Lan et al., 2007). Interestingly, sex differences in phenotypes of mouse mutants and human Kabuki syndrome provide evidence for significant functional divergence between X and Y homologues of this gene (Snell and Turner, 2018; Arnold, 2019), including lack of histone demethylase activity of the Y homologue (Shpargel et al., 2012). Substantial differences in nucleotide and predicted protein sequence identity (Supplementary Table 3.7) for X and Y homologues of *ZFX/ZFY* (95% and 89%) and *KDM6A* aka *UTX/UTY* (85% and 80%), and plots of organ weights vs. gametologue dosage (Supplementary Figure 3.10), further support functional divergence of *KDM6A* and *KDM6D* in the current dataset. Thus, for those gametologues and tissues that lack significant XX:XY dosage ratio bias (Supplementary Table 3.6), qualitative rather than quantitative differences may explain effects of XX and XY homologues on organ weights.

In conclusion, we have demonstrated that apart from *XIST* only gametologues consistently discriminate female and male transcriptomes across a range of fetal tissues that represent all three embryonic germ layers and the trophoctoderm. Correlation matrices (Supplementary Figure 3.11) indicate a tissue- and, likely, developmental stage-specific coordinated expression of gametologues that is consistent with their fundamental and central functions in the regulation of gene expression (Bellott et al., 2014; Seo et al., 2016). Combined with the finding that gametologue dosage of females and males accounts for a substantial proportion of the sex-specific phenotypic variation in organ weights, our data thus provide compelling evidence that gametologues play a major role in early female-male phenotypic differentiation.

References

- Arnold, A.P., 2019. Rethinking sex determination of non-gonadal tissues. *Curr top dev biol* [Online], 134, pp.289–315.
- Arnold, A.P., 2017. A general theory of sexual differentiation. *Journal of neuroscience research*, 95(1-2), pp.291–300.
- Bellott, D.W., Hughes, J.F., Skaletsky, H., Brown, L.G., Pyntikova, T., Cho, T.J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R.S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alfoldi, J., Rozen, S., Muzny, D.M., Warren, W.C., Gibbs, R.A., Wilson, R.K., and Page, D.C., 2014. Mammalian y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* [Online], 508(7497), pp.494–9.
- Gabory, A., Roseboom, T.J., Moore, T., Moore, L.G., and Junien, C., 2013. Placental contribution to the origins of sexual dimorphism in health and diseases: sex chromosomes and epigenetics. *Biol sex differ* [Online], 4(1), p.5.
- Gershoni, M. and Pietrokovski, S., 2017. The landscape of sex-differential transcriptome and its consequent selection in human adults. *Bmc biol* [Online], 15(1), p.7.
- Huret, J.L., Ahmad, M., Arsaban, M., Bernheim, A., Cigna, J., Desangles, F., Guignard, J.C., Jacquemot-Perbal, M.C., Labarussias, M., Leberre, V., Malo, A., Morel-Pair, C., Mossafa, H., Potier, J.C., Texier, G., Viguie, F., Yau Chun Wan-Senon, S., Zasadzinski, A., and Dessen, P., 2013. Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic acids res* [Online], 41(Database issue), pp.D920–4.
- Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., Iwase, S., Alpatov, R., Issaeva, I., Canaani, E., Roberts, T.M., Chang, H.Y., and Shi, Y., 2007. A histone h3 lysine 27 demethylase regulates animal posterior development. *Nature* [Online], 449(7163), pp.689–94.

- Liu, R., Low, W.Y., Tearle, R., Koren, S., Ghurye, J., Rhie, A., Phillippy, A.M., Rosen, B.D., Bickhart, D.M., Smith, T.P.L., Hiendleder, S., and Williams, J.L., 2019. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine x and y chromosomes. *Bmc genomics* [Online], 20(1), p.1000.
- Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D.M., Rosen, B.D., Kronenberg, Z.N., Kingan, S.B., Tseng, E., Thibaud-Nissen, F., Martin, F.J., Billis, K., Ghurye, J., Hastie, A.R., Lee, J., Pang, A.W.C., Heaton, M.P., Phillippy, A.M., Hiendleder, S., Smith, T.P.L., and Williams, J.L., 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in angus and brahman cattle. *Nat commun* [Online], 11(1), p.2071.
- Lowe, R., Gemma, C., Rakyan, V.K., and Holland, M.L., 2015. Sexually dimorphic gene expression emerges with embryonic genome activation and is dynamic throughout development. *Bmc genomics* [Online], 16, p.295.
- Sahakyan, A., Yang, Y., and Plath, K., 2018. The role of xist in x-chromosome dosage compensation. *Trends cell biol* [Online], 28(12), pp.999–1013.
- Seo, M., Caetano-Anolles, K., Rodriguez-Zas, S., Ka, S., Jeong, J.Y., Park, S., Kim, M.J., Nho, W.G., Cho, S., Kim, H., and Lee, H.J., 2016. Comprehensive identification of sexually dimorphic genes in diverse cattle tissues using rna-seq. *Bmc genomics* [Online], 17, p.81.
- Shpargel, K.B., Sengoku, T., Yokoyama, S., and Magnuson, T., 2012. Utx and uty demonstrate histone demethylase-independent function in mouse embryonic development. *Plos genet* [Online], 8(9), e1002964.
- Snell, D.M. and Turner, J.M.A., 2018. Sex chromosome effects on male-female differences in mammals. *Curr biol* [Online], 28(22), R1313–R1324.
- Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B.B., Castel, S.E., Karczewski, K.J., Aguet, F., Byrnes, A., Consortium, G.T., Laboratory, D.A., Coordinating Center

- Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E.B.I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U.o.C.S.C., Lapalainen, T., Regev, A., Ardlie, K.G., Hacohen, N., and MacArthur, D.G., 2017. Landscape of x chromosome inactivation across human tissues. *Nature* [Online], 550(7675), pp.244–248.
- Xiang, R., Ghanipoor-Samami, M., Johns, W.H., Eindorf, T., Rutley, D.L., Kruk, Z.A., Fitzsimmons, C.J., Thomsen, D.A., Roberts, C.T., Burns, B.M., Anderson, G.I., Greenwood, P.L., and Hiendleder, S., 2013. Maternal and paternal genomes differentially affect myofibre characteristics and muscle weights of bovine fetuses at midgestation. *Plos one* [Online], 8(1), e53402.
- Xiang, R., Lee, A.M., Eindorf, T., Javadmanesh, A., Ghanipoor-Samami, M., Gugger, M., Fitzsimmons, C.J., Kruk, Z.A., Pitchford, W.S., Leviton, A.J., Thomsen, D.A., Beckman, I., Anderson, G.I., Burns, B.M., Rutley, D.L., Xian, C.J., and Hiendleder, S., 2014. Widespread differential maternal and paternal genome effects on fetal bone phenotype at mid-gestation. *J bone miner res* [Online], 29(11), pp.2392–404.
- Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A., and Lusk, A.J., 2006. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome res* [Online], 16(8), pp.995–1004.

3.3 Methods

3.3.1 Animals, phenotypes and tissue sampling

Purebred and reciprocal cross Angus (*Bos taurus taurus*) and Brahman (*Bos taurus indicus*) conceptuses were generated and recovered at Day 153 of gestation (Anand-Ivell et al., 2011). Organ weights were recorded, and tissue samples were snap frozen in liquid nitrogen and stored at -80°C . Muscle weight was determined as described previously (Xiang et al., 2013). We sequenced the transcriptome of brain, liver, lung, skeletal muscle and placenta samples of 12 male and 12 female fetuses. All animal procedures were approved by the University of Adelaide Animal Ethics Committee (No. S-094-2005).

3.3.2 RNA preparation and sequencing

The total RNA from tissues was extracted from tissues using a RNeasy Plus Universal kit (Illumina, San Diego, CA) and ribosomal RNA was removed using a RiboZero Gold kit (Illumina, San Diego, CA). Sequencing libraries were prepared with a KAPA Stranded RNA-Seq Library Preparation Kit following the Illumina paired-end library preparation. Paired-end 100bp reads were generated using the Illumina Next-Seq 2000 sequencing system. Sequence data was evaluated with fastqc (Andrews et al., 2010). Four samples (Two male muscle samples, one male lung sample and one male placenta samples) were considered as low-quality samples, i.e., showed more than 400% technical and biological variation than other replicates in same group (Liu et al., 2015) and were discarded, leaving 60 female and 56 male samples with ≥ 70 million reads per sample for analysis.

3.3.3 Sex-specific expression analysis

For fetal samples, RNA-seq reads were mapped to the Brahman cattle reference genome (UOA_brahman_1) that has an additional non-PAR Y chromosome sequence from the Angus genome (UOA_angus_1) added to it (Liu et al., 2019; Low et al., 2020) using Hisat2 (Kim et al., 2015), then gene counts extracted using featurecounts (Liao et al., 2014) based on Ensembl annotation (v97). Samples were grouped by tissue and analysed using the limma package (Ritchie et al., 2015). To remove these low-expression genes that

may affect DEG detection sensitivity, sequence counts were log-transformed into counts per million (CPM) to standardise for differences in library size. Genes expressed in at least 3 female or male biological replicates at $CPM > 1$ were considered as expressed genes for further analysis. Sequence counts were normalised using the trimmed mean of M values (TMM) method (Robinson and Oshlack, 2010) to avoid bias due to different coverage. The sequence counts were log-transformed into counts per million (CPM) to standardise for differences in library size for removing non-expression genes. Genes expressed at CPM_{ge1} in at least 3 female or male biological replicates were considered as expressed genes and used for further analysis. Counts were normalised using the trimmed mean of M values (TMM) method (Robinson and Oshlack, 2010) to avoid bias due to different coverage. Gene counts combined with weights estimated from gene expression in each observation (i.e., each gene per sample) and weights from replicates between the two sexes were fitted to a linear model using `voomwithQualityWeights` function (Liu et al., 2015). Moderated t-statistics test was used to define differential expression levels between samples, and then ranked, based on false discovery rate (FDR) at an adjusted P-value < 0.05 (Ritchie et al., 2015).

For adult samples, raw RNA-Seq fastq files were obtained from GEO database GSE65125 (Seo et al., 2016). Four of these samples, one liver and three muscle samples, were identified as low-quality samples using same cut off (having more than 400% variability than other biological replicates in same group) as used for fetal samples. The remaining 36 samples were mapped to the same reference genome (Liu et al., 2019; Low et al., 2020) as described for the fetal tissues above. Subsequent analyses followed the same procedures and criteria as for the fetal samples to produce the list of differentially expressed genes.

3.3.4 Dosage compensation analysis

The dosage compensation of X chromosome gene expression was evaluated using modified Mann-Whitney U test proposed in Xiong et al. 2010 (Xiong et al., 2010) which was used to compare the expression levels of PAR genes/X-specific genes with autosomal genes. Briefly, for each sample, we multiplied the expression levels of all PAR/X-specific genes with a number called r , which is in the range of 0.5 to 10 and increased by 0.01. We then

compared the overall of these modified expression levels of PAR/X-specific genes with the original expression levels of autosomal genes using Mann-Whitney's U test. The value of $1/r$ which has largest p -value in Mann-Whitney's U test became our estimate of the PAR:A and X:A ratio.

3.3.5 Gametologue expression analysis

Of the 15 paired gametologues on the cattle sex chromosomes (Liu et al., 2019), 6 were not expressed and thus excluded from the analysis. The remaining 9 pairs of expressed gametologues were tested for balanced expression of the paired-sex-linked genes in both sexes. Sex-bias expression analysis in both fetal and adult tissues was also performed using counts of X gametologue genes in females and the combined counts of X and Y-gametologue genes in males. The pairwise correlation coefficients of expression ratios of the 9 paired expressed gametologues were calculated using the `cor.test` function in R.

3.3.6 Sex variation in organ weight explained by gametologues

The effect of sex on organ weight was estimated as the amount of variation explained (R^2) when fitting the factor sex alone in SAS Proc GLM (generalised linear model selection) in following model for each tissue:

$$\text{organ weight} = \alpha \cdot \text{sex}, \quad (3.1)$$

where α is coefficient.

Six differentially expressed (DE) gametologue pairs (*EIF2S3*, *KDM6A*, *SHROOM2*, *TXLNG*, *ZFX* and *ZRSR2*) were tested. To identify the important subset of gametologues for explaining the effect of sex on the weight of each organ, organ specific XX and combined XY expression values for the gametologue pairs of females and males were fitted as covariates in SAS Proc GLMSelect. The forward stepwise method was used, with adjusted R^2 as the selection and stop criterion.

The effect of each gametologue subset on organ weight was compared with the sex effect

using SAS Proc GLM in following model:

$$\begin{aligned} \text{organ weight} = & \alpha_1 \cdot \text{sex} + \alpha_2 \cdot \text{EIF2S3} + \alpha_3 \cdot \text{KDM6A} + \alpha_4 \cdot \text{SHROOM2} \\ & + \alpha_5 \cdot \text{TXLNG} + \alpha_6 \cdot \text{ZFX} + \alpha_7 \cdot \text{ZRSR2}, \end{aligned} \quad (3.2)$$

where gene symbols represent the expression value of given gene in each tissue.

Variation in organ weight was modelled as a function of the sex effect and the subset of gametologues for each tissue. Two models (M_1, M_2) were used: 1) fitting the sex effect before the gametologues, and; 2) fitting the gametologues before the sex effect. Type I mean squares (MS) for the sex effect were used to calculate the amount of the sex effect variation explained by the subset of gametologues (Equation 3.3): Sex effect explained by each organ specific gametologue subset which is

$$\frac{M_1 \cdot \text{sex MS} - M_2 \cdot \text{sex MS}}{M_1 \cdot \text{sex MS}}. \quad (3.3)$$

References

- Anand-Ivell, R., Hiendleder, S., Vinales, C., Martin, G.B., Fitzsimmons, C., Eurich, A., Hafen, B., and Ivell, R., 2011. *Insl3* in the ruminant: a powerful indicator of gender- and genetic-specific fetomaternal dialogue. *Plos one* [Online], 6(5), e19821.
- Andrews, S. et al., 2010. *Fastqc: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- Kim, D., Langmead, B., and Salzberg, S.L., 2015. Hisat: a fast spliced aligner with low memory requirements. *Nat methods* [Online], 12(4), pp.357–60.
- Liao, Y., Smyth, G.K., and Shi, W., 2014. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* [Online], 30(7), pp.923–30.
- Liu, R., Holik, A.Z., Su, S., Jansz, N., Chen, K., Leong, H.S., Blewitt, M.E., Asselin-Labat, M.L., Smyth, G.K., and Ritchie, M.E., 2015. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic acids res* [Online], 43(15), e97.
- Liu, R., Low, W.Y., Tearle, R., Koren, S., Ghurye, J., Rhie, A., Phillippy, A.M., Rosen, B.D., Bickhart, D.M., Smith, T.P.L., Hiendleder, S., and Williams, J.L., 2019. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine x and y chromosomes. *Bmc genomics* [Online], 20(1), p.1000.
- Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D.M., Rosen, B.D., Kronenberg, Z.N., Kingan, S.B., Tseng, E., Thibaud-Nissen, F., Martin, F.J., Billis, K., Ghurye, J., Hastie, A.R., Lee, J., Pang, A.W.C., Heaton, M.P., Phillippy, A.M., Hiendleder, S., Smith, T.P.L., and Williams, J.L., 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in angus and brahman cattle. *Nat commun* [Online], 11(1), p.2071.

- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K., 2015. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids res* [Online], 43(7), e47.
- Robinson, M.D. and Oshlack, A., 2010. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biol* [Online], 11(3), R25.
- Seo, M., Caetano-Anolles, K., Rodriguez-Zas, S., Ka, S., Jeong, J.Y., Park, S., Kim, M.J., Nho, W.G., Cho, S., Kim, H., and Lee, H.J., 2016. Comprehensive identification of sexually dimorphic genes in diverse cattle tissues using rna-seq. *Bmc genomics* [Online], 17, p.81.
- Xiang, R., Ghanipoor-Samami, M., Johns, W.H., Eindorf, T., Rutley, D.L., Kruk, Z.A., Fitzsimmons, C.J., Thomsen, D.A., Roberts, C.T., Burns, B.M., Anderson, G.I., Greenwood, P.L., and Hiendleder, S., 2013. Maternal and paternal genomes differentially affect myofibre characteristics and muscle weights of bovine fetuses at midgestation. *Plos one* [Online], 8(1), e53402.
- Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X., 2010. Rna sequencing shows no dosage compensation of the active x-chromosome. *Nat genet* [Online], 42(12), pp.1043–7.

3.4 Supplementary Figures

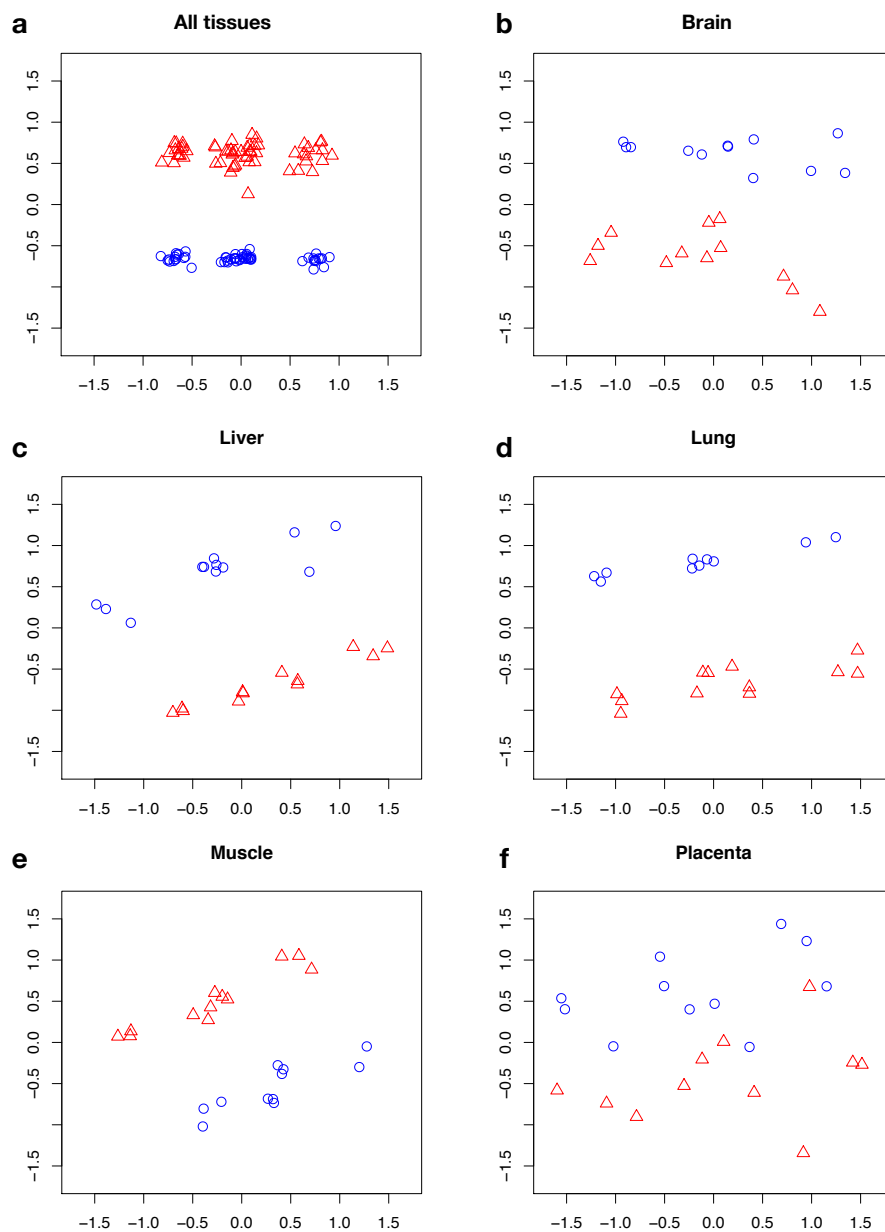


Figure 3.3: Multi-dimensional scaling plots of male-female difference on a two-dimensional scatterplot.

Male samples in blue, female samples in red, the X and Y axes are in log₂ fold changes of gene expression. Distances on the plot can be interpreted as log₂-fold-change between the samples for the genes that distinguish those samples. a) Combined samples from all five tissues, b-e) brain, liver, lung, muscle and placenta.

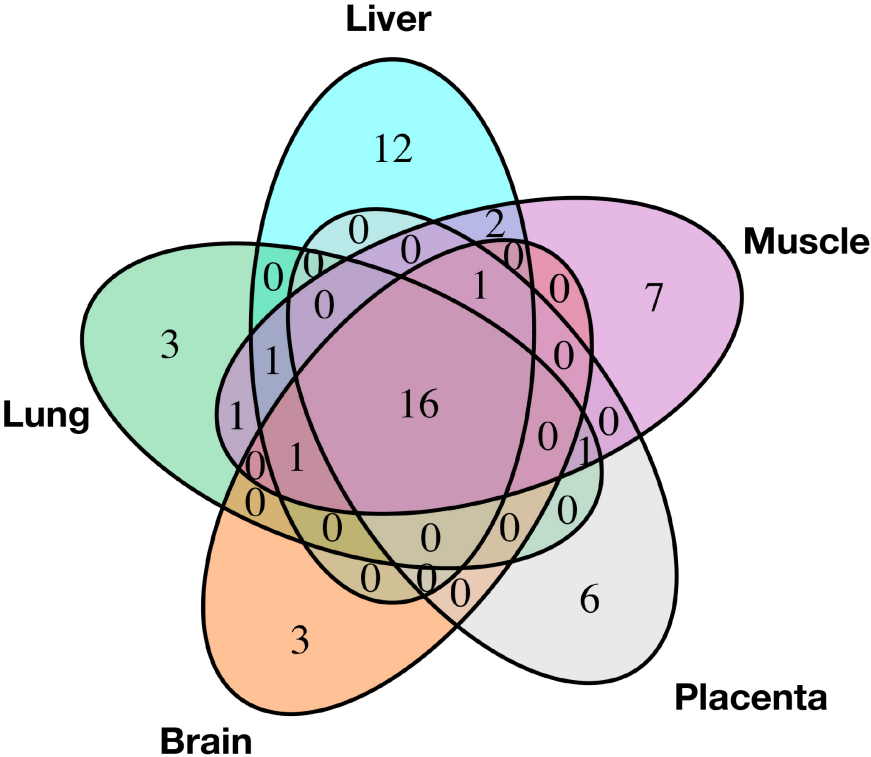


Figure 3.4: Venn diagram with numbers of differentially expressed genes (FDR < 0.05) between males and females in five fetal tissues.

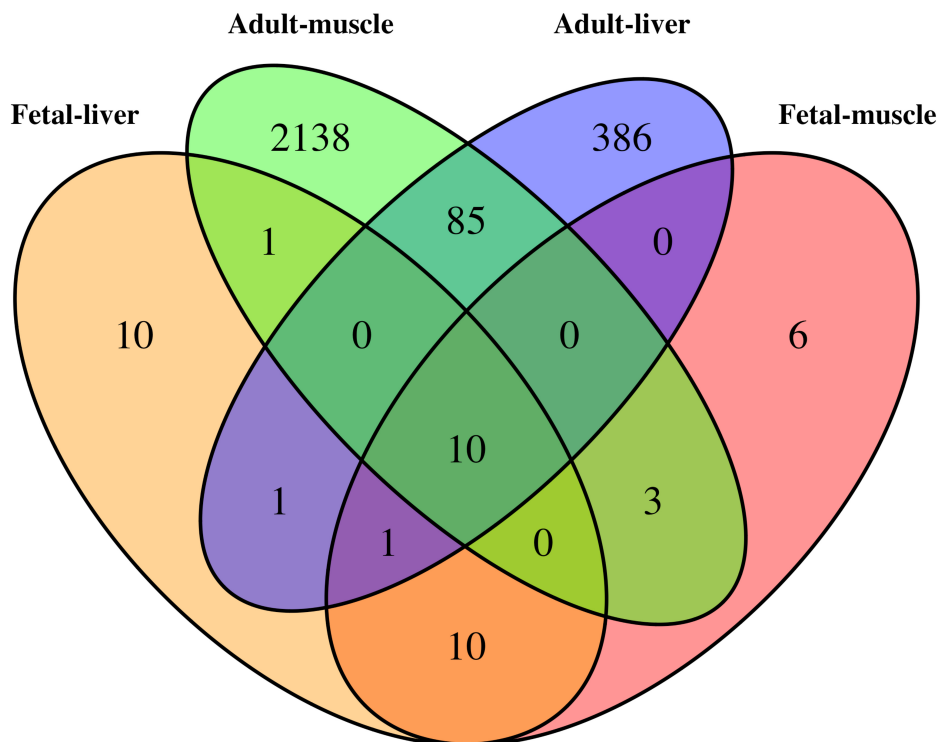


Figure 3.5: Venn diagram showing the overlap between differentially expressed (DE) genes (FDR<0.05) between males and females in fetal and adult tissues.

Total DE genes in fetal liver and muscle is 33 and 30 respectively as compared with 483 and 2,237 in adult liver and muscle, respectively. The 10 DE genes shared between fetal and adult tissues include one lncRNA (*XIST*) and 9 Y-linked gametologues (*OFD1Y*, *USP9Y*, *EIF2S3Y*, *ZRSR2Y*, *UTY*, *DDX3Y*, *SHROOM2Y*, *ZFY*, *TXLNGY*).

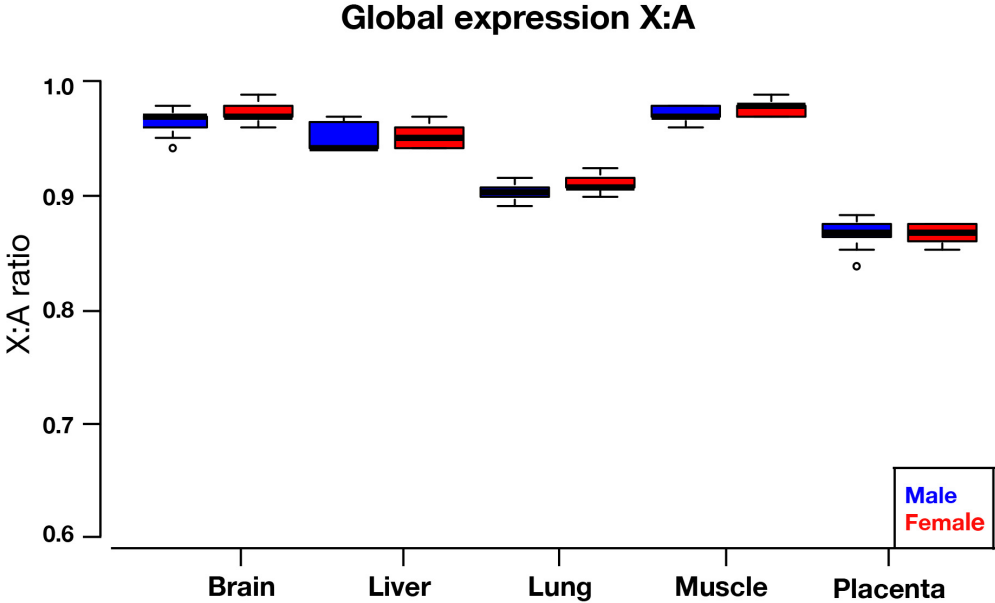


Figure 3.6: Ratio of the median expression levels of X-specific genes and autosomal genes of female (red) and male (blue) samples.

There are no significant differences between female and male samples.

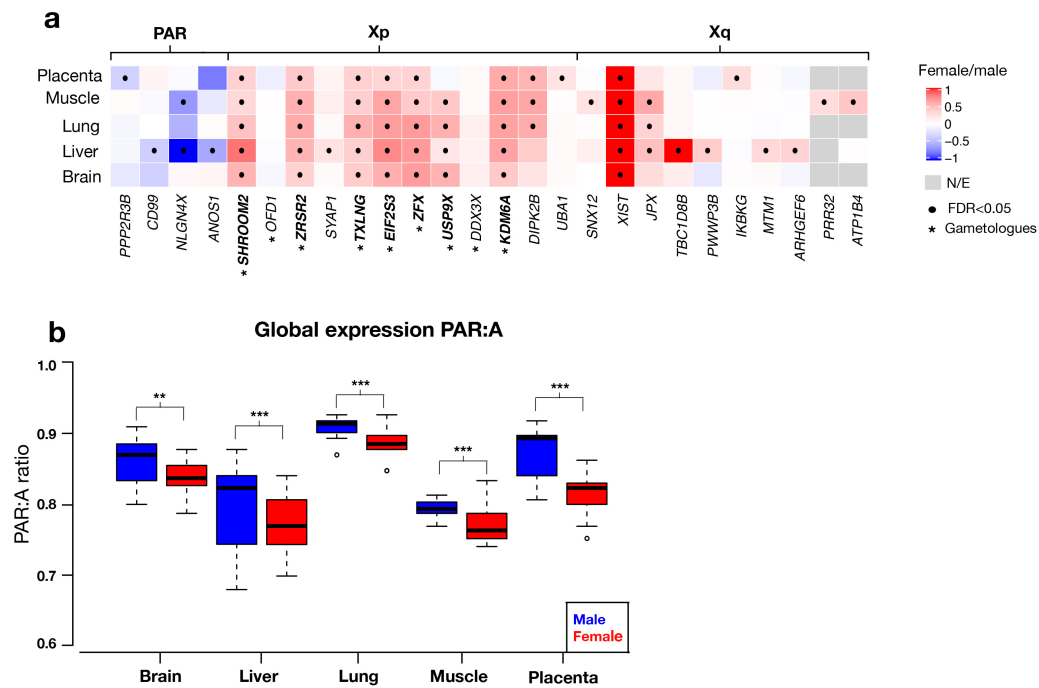


Figure 3.7: Heatmap of differentially expressed (DE) genes between females and males and ratios of the median expression levels of X chromosome genes and autosomal genes in female and male fetal tissues.

a) Heatmap representation of the female and male expression difference in 26 X-linked genes located on PAR, Xp (the short arm of X chromosome) and Xq (the long arm of X chromosome) across 5 tissues. The colour scale displays the direction of sex-bias with red colour indicating higher female expression. Significantly DE genes are indicated by a black dot. Non-expressed genes in the sex-bias analysis are in grey. Gene names of X-Y chromosome paired gametologues are in bold and labelled with an asterisk. b) Ratio of the median expression levels of pseudoautosomal region (PAR) genes ($n=28$) and all autosomal genes (brain $n = 16226$, liver $n = 16371$, lung $n = 16259$, muscle $n = 15818$, placenta $n = 15761$) of female (red) and male (blue) samples. Asterisks indicate significant differences between female and male samples (paired Wilcoxon rank-sum test, $P < 0.05$).

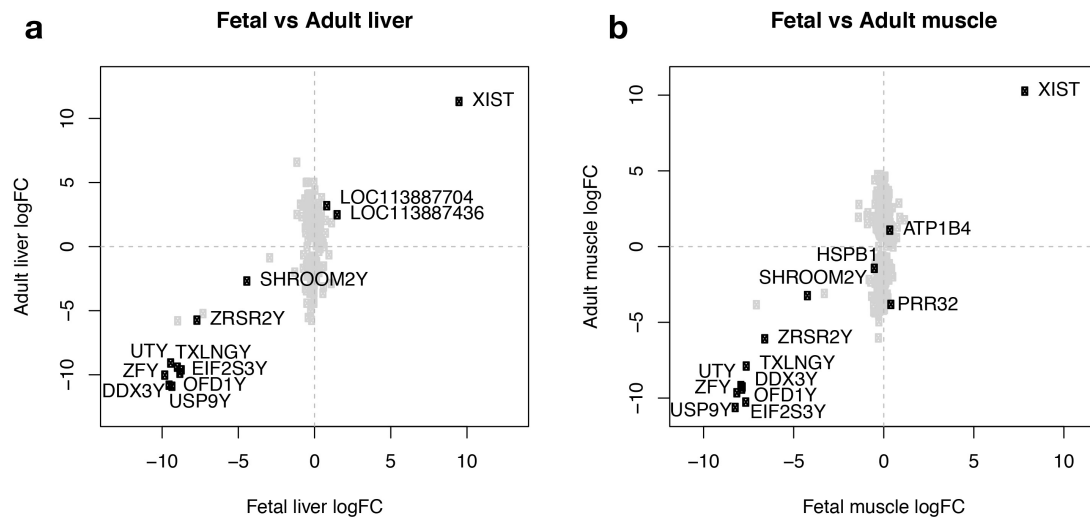


Figure 3.8: Comparison of differentially expressed (DE) genes between females and males in fetal and adult tissues.

Both plots include all DE genes discovered in fetal and adult tissues (FDR < 0.05). DE genes overlapping in both developmental stages are highlighted in black and identified by gene symbols. Non-overlapping DE genes are in grey. X-axis is the log₂ fold changes in fetal tissues. Y-axis is the log₂ fold changes in adult tissues. a) The average log₂ fold changes of female-male DE genes in fetal and adult liver. b) The average log₂ fold changes of female-male DE genes in fetal and adult muscle.

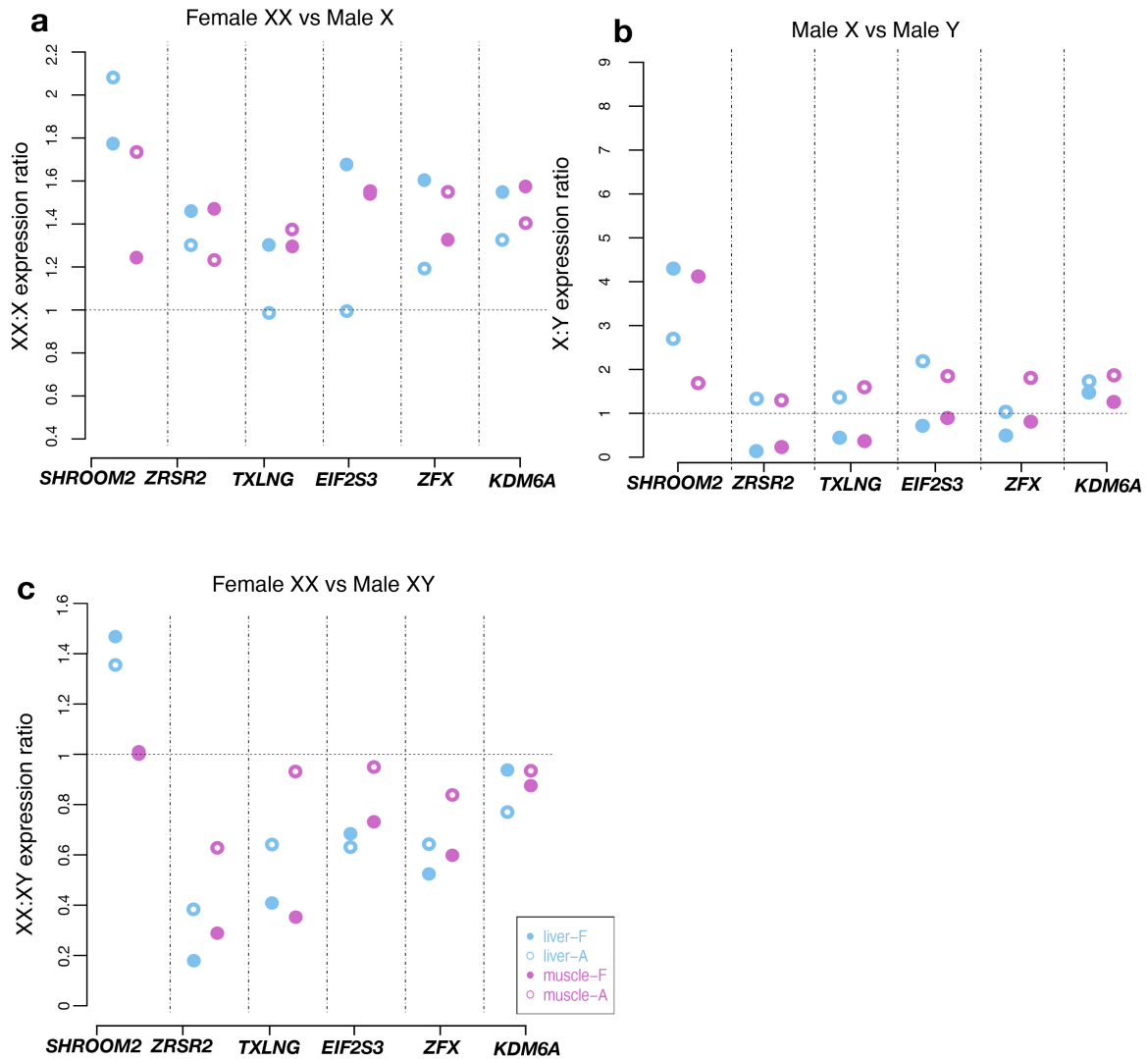


Figure 3.9: Comparison of mean XX:X, X:Y and XX:combined XY expression ratios of gametologues in two fetal (F) and adult (A) tissues.

a) Ratio of X-chromosome gametologues in female and male liver (blue) and muscle (purple). Gametologue expression ratios for fetal tissues are shown as filled circles, ratios for adult tissues as open circles. The dashed line represents equal expression ratio for female and male. b) Ratio of mean expression levels of X and Y gametologues in males. The dashed line represents equal expression ratio for male X and male Y. c) Ratio of mean expression levels of X-gametologues in females and combined expression of X- and Y-gametologues in males. The dashed line represents equal expression ratio for female XX and combined male XY.

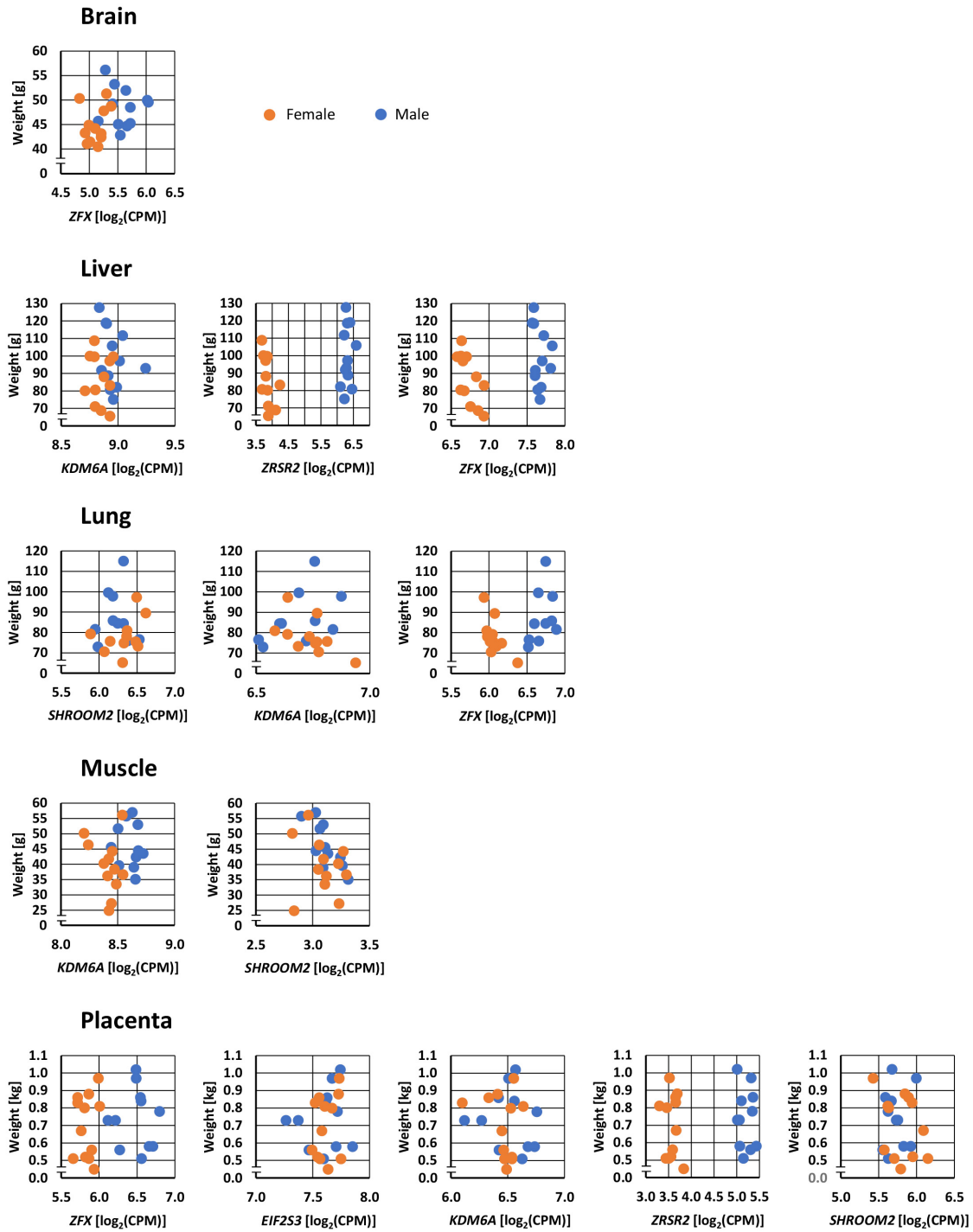


Figure 3.10: Organ weights of females and males and expression levels of gametologues that explain variation captured by the factor ‘sex’ in linear models.

Expression level is depicted as \log_2 of gametologue transcript counts per million (CPM).

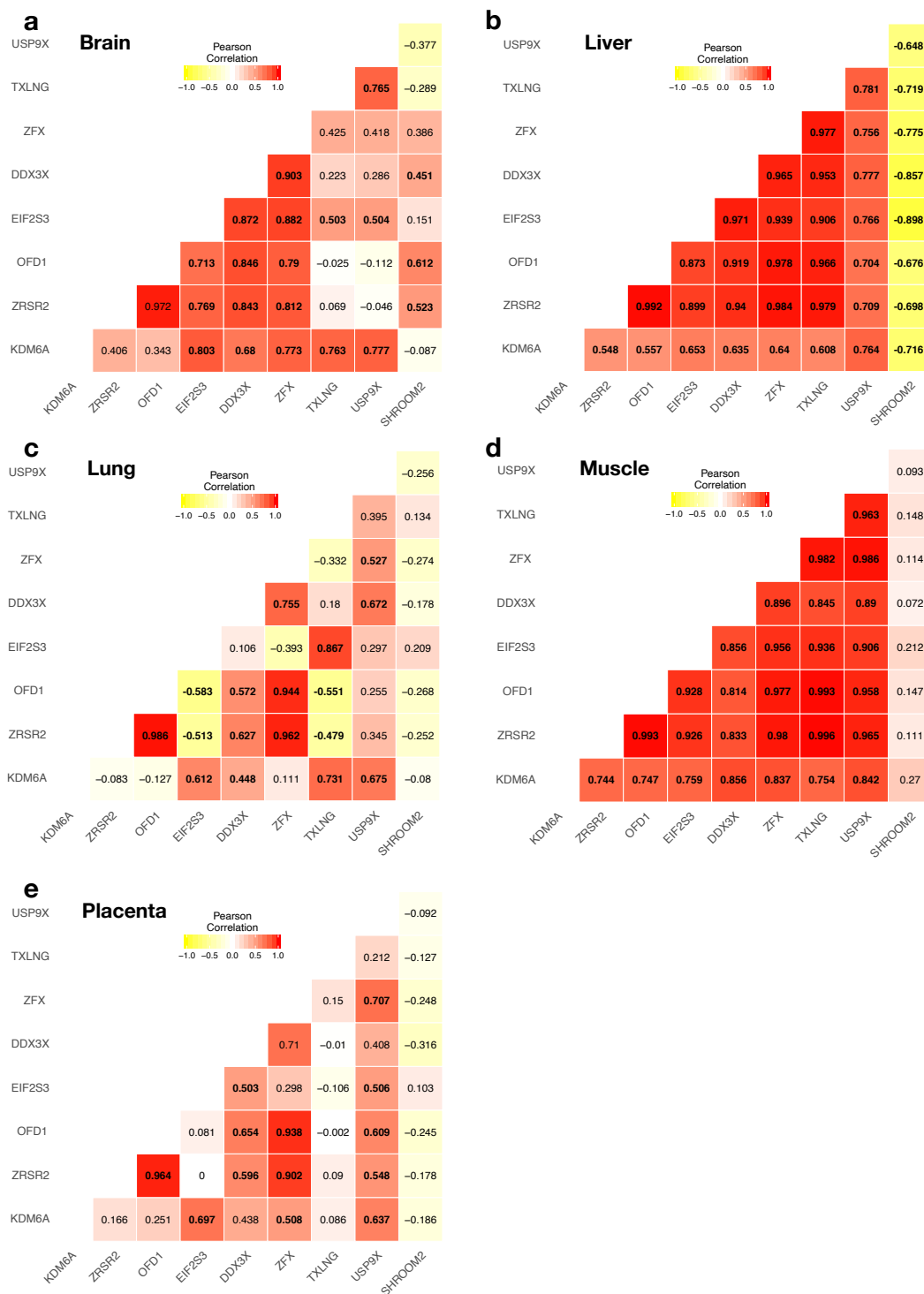


Figure 3.11: Correlations between expression levels of gametologues within tissues.

a-e) Heatmap of correlation matrix between 9 gametologues brain, liver, lung, muscle and placenta. For female samples, expression value is from X-gametologues. For male samples, expression value is the combined expression of X and Y gametologues. Significant correlations (p -value<0.05) are in bold.

3.5 Supplementary Tables

Table 3.1: Number of expressed genes by chromosome and fetal tissue.
Y* chromosome here refers to only the non-PAR Y sequence.

Chromosome	Brain	Liver	Lung	Muscle	Placenta
1	654	692	674	657	640
2	702	688	700	700	687
3	941	961	971	940	934
4	537	560	551	541	532
5	890	918	919	871	885
6	458	462	447	457	427
7	939	938	922	890	909
8	552	548	547	551	522
9	407	411	392	404	389
10	708	700	694	693	699
11	791	769	782	760	754
12	327	312	303	303	290
13	638	628	640	601	620
14	402	384	392	386	370
15	509	519	505	491	492
16	536	552	534	525	511
17	517	504	509	486	481
18	925	933	939	877	903
19	962	958	985	934	941
20	253	260	257	261	256
21	444	453	448	431	425
22	473	471	465	453	450
23	476	519	492	486	479
24	243	251	243	231	231
25	617	602	627	597	602
26	304	309	293	294	305
27	181	186	180	181	185
28	243	246	241	231	234
29	466	461	468	439	486
X	632	629	649	630	594
Y*	12	13	13	13	13
All chromosomes	16739	16837	16782	16314	16246

Table 3.2: Nucleotide and protein sequence identity of X- and Y-chromosome gametologues.

Nucleotide	X	Y	Coverage	Similarity
KDM6A/UTY	ENSBTAT00000063826	ENSBIXT00000053168	98%	84.69%
ZRSR2/ZRSR2Y	XM_003588204.5	ENSBIXT00000053041	70%	83.90%
EIF2S3/EIF2S3Y	ENSBTAT00000019064	ENSBIXT00000053260	100%	90.86%
ZFX/ZFY	ENSBTAT00000010165	ENSBIXT00000053251	94%	94.94%
TXLNG/TXLGY	ENSBTAT00000002316	ENSBIXT00000053401	49%	83.40%
SHROOM2/SHROOM2Y	ENSBTAT00000000256	ENSBIXT00000053233	30%	96.74%
USP9X/USP9Y	ENSBTAT00000050390	ENSBIXT00000053035	100%	89.54%
OFD1/OFD1Y	ENSBTAT00000005850	ENSBIXT00000053027	81%	86.39%
DDX3X/DDX3Y	ENSBTAT00000050399	ENSBIXT00000053216	100%	87.17%
Protein	X	Y	Coverage	Identity
KDM6A/UTY	ENSBTAP00000056061	ENSBIXP00000000430	99%	80.08%
ZRSR2/ZRSR2Y	XP_003588252.2	ENSBIXP00000026597	98%	77.66%
EIF2S3/EIF2S3Y	ENSBTAP00000019064	ENSBIXP00000000372	99%	97.44%
ZFX/ZFY	ENSBTAP00000010165	ENSBIXP00000000381	100%	89.39%
TXLNG/TXLGY	ENSBTAP00000002316	ENSBIXP00000000309	67%	81.75%
SHROOM2/SHROOM2Y	ENSBTAP00000000256	ENSBIXP00000026523	43%	94.50%
USP9X/USP9Y	ENSBTAP00000047095	ENSBIXP00000026598	100%	92.52%
OFD1/OFD1Y	ENSBTAP00000005850	ENSBIXP00000026600	91%	72.03%
DDX3X/DDX3Y	ENSBTAP00000047104	ENSBIXP00000026528	100%	88.70%

Table 3.3: Differentially expressed genes between females and males in five tissues (FDR <0.05).

Chr	Start	End	Symbol	gene_type	Placenta	Muscle	Brain	Lung	Liver
X	67152847	67190784	XIST	lncRNA	Placenta	Muscle	Brain	Lung	Liver
Y	7101173	7226681	USP9Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	8056568	8104026	EIF2S3Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	7649743	7686255	DDX3Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	7305926	7361308	ZRSR2Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	10463946	10554886	TXLNGY	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	7419104	7520643	UTY	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	7966867	8039794	ZFY	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	7695294	7730193	SHROOM2Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
Y	6953048	7017533	OFD1Y	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	43956970	44136596	KDM6A	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	64545574	64556153	LOC113887436	lncRNA	Placenta	Muscle	Brain	NA	Liver
X	44168352	44214281	DIPK2B	protein_coding	Placenta	Muscle	NA	Lung	NA
X	13448650	13472008	ZRSR2	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	21901312	21916406	EIF2S3	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	21971657	22012094	ZFX	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	107688248	107713140	IKBKG	protein_coding	Placenta	NA	NA	NA	NA
X	14425649	14469182	TXLNG	protein_coding	Placenta	Muscle	Brain	Lung	Liver
5	19752901	19756864	LOC113892437	lncRNA	Placenta	NA	NA	NA	NA
5	45201656	45220300	FOXRED2	protein_coding	Placenta	NA	NA	NA	NA
3	110331467	110340607	NCDN	protein_coding	Placenta	NA	NA	NA	NA
X	47533	88481	PPP2R3B	protein_coding	Placenta	NA	NA	NA	NA
X	6861167	7002254	SHROOM2	protein_coding	Placenta	Muscle	Brain	Lung	Liver
X	58504862	58526800	UBA1	protein_coding	Placenta	NA	NA	NA	NA
X	40335983	40452913	USP9X	protein_coding	NA	Muscle	Brain	Lung	Liver
X	21924125	21925702	LOC113887341	lncRNA	NA	Muscle	NA	Lung	NA
X	67228381	67416605	LOC113886777	lncRNA	NA	Muscle	NA	Lung	Liver
X	141117939	141165044	ATP1B4	protein_coding	NA	Muscle	NA	NA	NA
19	25241519	25258332	SPATA22	protein_coding	NA	Muscle	NA	NA	Liver
X	64592196	64655269	SNX12	protein_coding	NA	Muscle	NA	NA	NA
25	8194683	8196390	HSPB1	protein_coding	NA	Muscle	NA	NA	NA
X	134912621	134914691	PRR32	protein_coding	NA	Muscle	NA	NA	NA
12	72138600	72154359	LOC113902493	lncRNA	NA	Muscle	NA	NA	NA
21	25634171	25671362	MINAR1	protein_coding	NA	Muscle	NA	NA	NA
4	109552769	109813267	CDK6	protein_coding	NA	Muscle	NA	NA	NA
X	3433026	3826465	NLGN4X	protein_coding	NA	Muscle	NA	NA	Liver
Y	6911360	6916060	AMELY	protein_coding	NA	NA	Brain	NA	NA
1	147316800	147356339	LOC113896378	lncRNA	NA	NA	Brain	NA	NA
11	36516335	36695724	ACYP2	protein_coding	NA	NA	Brain	NA	NA
X	7048559	7058641	LOC113887319	protein_coding	NA	NA	NA	Lung	NA

Continued on next page

Table 3.3 – continued from previous page

Chr	Start	End	Symbol	gene_type	Placenta	Muscle	Brain	Lung	Liver
18	17339117	17346774	LGALS4	protein_coding	NA	NA	NA	Lung	NA
18	3769595	3774548	SBK2	protein_coding	NA	NA	NA	Lung	NA
X	89284004	89361758	TBC1D8B	protein_coding	NA	NA	NA	NA	Liver
X	7008965	7014078	LOC113887704	lncRNA	NA	NA	NA	NA	Liver
X	1295715	1326007	CD99	protein_coding	NA	NA	NA	NA	Liver
19	28779211	28799925	ALOXE3	protein_coding	NA	NA	NA	NA	Liver
X	5704955	5916854	ANOS1	protein_coding	NA	NA	NA	NA	Liver
X	93021009	93084387	PWWP3B	protein_coding	NA	NA	NA	NA	Liver
17	27806462	27809658	LOC113907639	lncRNA	NA	NA	NA	NA	Liver
9	60816375	60873221	RRAGD	protein_coding	NA	NA	NA	NA	Liver
X	125543411	125660731	ARHGEF6	protein_coding	NA	NA	NA	NA	Liver
X	112832096	112924518	MTM1	protein_coding	NA	NA	NA	NA	Liver
15	29586729	29639150	SLCO2B1	protein_coding	NA	NA	NA	NA	Liver
X	14370937	14386915	SYAP1	protein_coding	NA	NA	NA	NA	Liver

Table 3.4: Summary statistics of phenotype data of fetuses and of expression values for XX and combined XY gametologues.

Organ	Number of Weights [g]	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
Brain		24	46.7	4.2	40.5	56.2
Liver		24	93.1	16.6	65.5	127.6
Lung		24	81.6	13.8	50.3	115
Muscle		24	42.1	9	24.9	57
Placenta		24	728	176	450	1020

Organ-specific expression values for XX and combined XY gametologues [$\log_2(\text{CPM})$]¹

Brain						
<i>EIF2S3</i>		24	5.79	0.2	5.37	6.27
<i>KDM6A</i>		24	4.67	0.64	3.84	5.54
<i>SHROOM2</i>		24	4.96	1.15	3.56	6.27
<i>ZFX</i>		24	7.55	0.21	7.24	8.04
<i>ZRSR2</i>		24	5.35	0.33	4.83	6.04
Liver						
<i>EIF2S3</i>		24	8.9	0.11	8.71	9.24
<i>KDM6A</i>		24	5.09	1.26	3.68	6.59
<i>SHROOM2</i>		24	5.67	0.82	4.71	6.68
<i>ZFX</i>		24	7.18	0.4	6.63	7.89
<i>ZRSR2</i>		24	7.2	0.49	6.58	7.83
Lung						
<i>EIF2S3</i>		22	6.71	0.11	6.51	6.94
<i>KDM6A</i>		22	5.37	0.7	4.54	6.23
<i>SHROOM2</i>		22	5.8	1.07	4.67	7.12
<i>ZFX</i>		22	8.3	0.12	8.08	8.52
<i>ZRSR2</i>		22	6.35	0.35	5.93	6.89
Muscle						
<i>EIF2S3</i>		23	8.51	0.14	8.2	8.72
<i>KDM6A</i>		23	4.97	0.91	4	6.05
<i>SHROOM2</i>		23	5.59	0.91	4.59	6.64
<i>ZFX</i>		23	7.15	0.14	6.89	7.33
<i>ZRSR2</i>		23	7.51	0.39	6.97	8.02
Placenta						
<i>EIF2S3</i>		23	6.49	0.17	6.1	6.76
<i>KDM6A</i>		23	4.35	0.84	3.3	5.43
<i>SHROOM2</i>		23	4.69	0.87	3.53	5.88
<i>ZFX</i>		23	8.53	0.2	8.09	8.84
<i>ZRSR2</i>		23	6.15	0.37	5.66	6.8

¹for comparison of sex differences by organ see supplementary Figure 3.10

Table 3.5: Mean organ weights and variation explained by the sex effect.

Mean Weights [g]	Brain	Liver	Lung	Muscle	Placenta
Male	48.5	99.3	87.5	44.5	759
Female	44.9	86.9	75.8	39.6	698
Standard Error of the Mean [g]	1.1	4.5	3.7	2.6	51
Variation Explained by Sex [%]	19.3	14.6	18.4	7.8	3.2
Significance of Sex Effect [P]	0.032	0.065	0.036	0.188	0.403

Table 3.6: SAS Proc GLMSelect results for gametologue subset selection by organ.

Brain, forward Selection Summary				
Step	Effect	Number	R ²	
		Effects In	Model	Adjusted
0	Intercept	1	0	0
1	<i>ZFX</i>	2	0.2161	0.1805
Stop Details				
Entry Candidate	<i>TXLNG</i>			0.1689
Liver, forward Selection Summary				
Step	Effect	Number	R ²	
		Effects In	Model	Adjusted
0	Intercept	1	0	0
1	<i>KDM6A</i>	2	0.1318	0.0923
2	<i>ZRSR2</i>	3	0.3042	0.2379
3	<i>ZFX</i>	4	0.4319	0.3467
Stop Details				
Entry Candidate	<i>EIF2S3</i>			0.3396
Lung, forward Selection Summary				
Step	Effect	Number	R ²	
		Effects In	Model	Adjusted
0	Intercept	1	0	0
1	<i>SHROOM2</i>	2	0.1841	0.1433
2	<i>KDM6A</i>	3	0.2654	0.1881
3	<i>ZFX</i>	4	0.3399	0.2298
Stop Details				
Entry Candidate	<i>EIF2S3</i>			0.1886
Muscle, forward Selection Summary				
Step	Effect	Number	R ²	
		Effects In	Model	Adjusted
0	Intercept	1	0	0
1	<i>KDM6A</i>	2	0.1335	0.0922
2	<i>SHROOM2</i>	3	0.2001	0.1201
Stop Details				
Entry Candidate	<i>EIF2S3</i>			0.1081
Placenta, forward Selection Summary				
Step	Effect	Number	R ²	
		Effects In	Model	Adjusted
0	Intercept	1	0	0
1	<i>ZFX</i>	2	0.2414	0.2052
2	<i>EIF2S3</i>	3	0.4089	0.3498
3	<i>KDM6A</i>	4	0.491	0.4106
4	<i>ZRSR2</i>	5	0.5512	0.4514
5	<i>SHROOM2</i>	6	0.5819	0.4589
Stop Details				
Entry Candidate	<i>TXLNG</i>			0.4252

Table 3.7: Proportion of Sex Effect Variation in Organ Weight Explained by Gametologue Subsets.

	Brain	Liver	Lung	Muscle	Placenta
Variation Explained [%] ¹	96	69	58	34	18
Sex Type I SS ²	77.44	922.3	727.8	239.5	11272
Sex Type III SS	3.25	286.2	306	158.8	9240

¹ Variation Explained = (Type I SS – Type III SS)/Type I SS

² Sex effect fitted as the first effect in the linear model

4 Different cattle breeds show distinctive gene expression patterns, including imprinting signatures in reciprocal crosses

4.1 Abstract

There are two subspecies of cattle, *Bos taurus taurus* and *Bos taurus indicus* which have arisen from independent domestication events resulting in a large phenotypic and genetic differences. Some phenotypic differences between indicine and taurine breeds emerge during fetal development and are reflected in birth outcomes, including birth weight. We used an RNA-seq approach to explore expression profiles in the placenta and four somatic tissues of fetuses at mid-gestation from two cattle breeds, Angus and Brahman, and their reciprocal crosses.

We identified a large number of genes that showed significant breed difference in expression in each tissue. These genes were found to participate in pathways related to tissue-specific function. There were 110 differentially expressed genes (DEGs) between Angus and Brahman in all tissues which were related to functions including immune response and stress response. The liver was the only tissue with a substantial number of DEGs between reciprocal crosses, of which 310 overlapped with genes that were DE between purebred groups. Pathway analysis showed these overlapping DEGs were significantly enriched in metabolic processes. The DEG between the purebred groups and in the reciprocal crosses showed an additive expression pattern, where both paternal and maternal genomes contributed to the gene expression levels. Only 5% of DEGs in each tissue showed a parent of origin driven expression, Angus or Brahman, and showed both maternal and paternal dominant effects.

These data identify candidate genes potentially driving tissue-specific breed differences, and also provide biological insight into parental genome effects underlying phenotypic differences in cattle fetal development.

4.2 Introduction

There are large phenotypic and genetic differences among cattle breeds, and in particular between indicine and taurine breeds (Consortium et al., 2009, Hayes and Daetwyler, 2019). The taurine and indicine subspecies of cattle have arisen from independent domestication events resulting in a high degree of genetic divergence (Pitt et al., 2019). Indicine cattle are

more tolerant of hot, humid environments, and tick challenge and hence are better adapted to survive in tropical areas (Zeng et al., 2019). However, the productivity of indicine cattle is lower than taurine cattle, therefore, crossbreeding has been used to harness the positive traits of the two types to improve the performance of cattle in tropical environments (Menéndez-Buxadera et al., 2016). Genes such as *MSRB3* and *PLAG1* are involved in energy and muscle metabolism which affect weight and body condition of indicine cattle more than temperate cattle (Porto-Neto et al., 2014). However, the genetic factors involved in adaptation to tropical conditions remain largely unknown.

The phenotypic differences between indicine and taurine breeds emerge during fetal development (Mao et al., 2008) and are reflected in birth outcomes, including birth weight. Fetal growth rate accelerates after mid-gestation (~day 150) (Krog et al., 2018), and fetal tissue phenotypes, such as fetal bone and muscle weight, differ between Brahman (indicine), and Angus (taurine) fetus, and in their reciprocal crosses at mid-gestation (Xiang et al., 2013; Xiang et al., 2014). Maternally inherited genes have been shown to contribute disproportionately to myofibre development and muscle mass (Xiang et al., 2013; Xiang et al., 2014).

With advances in genome sequencing technology, transcriptome complexity and dynamics can now be explored in detail. Studies of the gene expression of adult bovine tissues, including muscle (Berton et al., 2016), liver (Alexandre et al., 2015; Mukiibi et al., 2018), mammary gland (Cui et al., 2014) and adipose tissue (Sheng et al., 2014) have identified genetic factors that contribute to the differences in feed efficiency, milk composition and deposition of intramuscular fat. However, there is little information available on differences in gene expression between breeds during fetal development. A comparison of gene expression between taurine and indicine breeds may provide biological insights into the origin of their phenotypic differences.

In this study, the transcriptome of the fetal placenta and four somatic tissues (brain, liver, lung, skeletal muscle (*M. quadriceps femoris*)), at mid-gestation from two cattle breeds, Angus and Brahman, and their reciprocal crosses was investigated. The differentially expressed genes detected between the breeds and between the reciprocal crosses are

related to tissue-specific functions and may explain some of the phenotypic differences observed.

4.3 Material and Methods

4.3.1 Animals and sample collection

All animal experiments and procedures described in this study were approved by the University of Adelaide Animal Ethics Committee (No. S-094-2005 and S-094-2005A). The animals and semen used were pure bred Angus (*Bos taurus taurus*) and Brahman (*Bos taurus indicus*) cattle, subsequently referred to as Bt and Bi respectively. Pure Bt and Bi females of approximately 16-20 months of age were maintained on pasture supplemented with silage. Cows were inseminated with semen of pure bred Bt or Bi sires and pregnancy tested by ultrasound scanning. Cows and fetuses were humanely sacrificed by stunning and exsanguination at day 153 ± 1 of gestation. Fetuses were dissected, tissues snap-frozen in liquid nitrogen and stored at -80°C as previously described (Xiang et al., 2013). The five tissues used in this study were brain, liver, lung, muscle and placenta from 3 male and 3 female fetuses, from each of the 4 genetic types (Bt×Bt, Bi×Bt, Bt×Bi, Bi×Bi; paternal genome listed first), giving a total of 24 samples per tissue.

4.3.2 RNA isolation, library preparation and sequencing

Total RNA was isolated from tissues using RNeasy Plus universal kit and ribosomal RNAs were removed using RiboZero Gold kit, in accordance with the manufacturer's recommendations (Illumina, San Diego, CA). Sequencing libraries were prepared with a KAPA Stranded RNA-Seq Library Preparation Kit following the Illumina paired-end library preparation protocol (Illumina, San Diego, CA). Paired-end (PE) 100bp sequence reads were produced on an Illumina Next-Seq 2000 platform.

4.3.3 Data analysis

FastQC (Andrew, 2010) was used to assess read quality and adaptor sequences were removed using cutadapt (Martin, 2011). A modified Bovine Brahman reference genome,

consisting of the autosomes and X chromosome from UOA_Brahman_1 and the non-PAR Y chromosome from UOA_Angus_1 (GCA_003369695.2; GCA_003369685.2) was used. Reads were aligned to this reference using `hisat2` with default setting (Kim et al., 2015). The number of unique mapped reads for each gene was calculated using feature counts from the `Rsubread` package (Liao et al., 2019), using gene definitions from Refseq and Ensembl annotation V97. Genes with a count per million reads (CPM) below 0.5 were excluded. Multiscale-dimensional (MDS) plots were created using `plotMDS` from the `limma` R package. The expression of genes was normalised across the libraries by the Trimmed Mean of M-values (TMM) (Robinson and Oshlack, 2010), and variation among samples resulting from differences between sequence runs was standardised using `RemoveBatchEffect` in the `limma` package. After down-weighting replicates with high variation, differentially expressed genes (DEGs) between breed groups with a false discovery rate (FDR) <0.05 were identified using the `limma-voom` R package (Law et al., 2014; Liu et al., 2015). Unnamed protein-coding genes were annotated using BLASTN with the nucleotide collection nr/nt (Camacho et al., 2009). Only genes with more than 90% identity to an annotated gene were accepted as that gene.

4.3.4 Functional analysis of DEGs

To allow functional analysis of DEGs, cattle gene IDs were converted to their homologous (human) using `BioMart` R packages (Durinck et al., 2009). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) enrichment analyses of DEGs were performed using the `limma` R package (Ritchie et al., 2015). GO terms for molecular functions (MF), biological processes (BP) and cellular components (CC) were interrogated. Fisher's exact tests were carried out and an adjusted p -value calculated using the Benjamini-Hochberg Procedure for multiple tests (FDR). Genes with an adjusted p -value <0.05 were considered to be differentially expressed (DE). GESA software was used to define and plot the pathway networks for DEGs.

4.3.5 Identification of Brahman/Angus gene expression pattern in crossbred groups

To compare the gene expression patterns between genotypes, the average expression of the 4 genetic groups was initially calculated for the genetic groups Bt×Bt, Bi×Bt, Bt×Bi, and Bi×Bi (subsequently identified as 1, 2, 3 and 4 respectively). For each gene, the absolute difference in average expression for each of the two groups being compare was calculated. For example, for given gene, diff14 denote the absolute difference between the average expression of group Bt×Bt and the average expression of group Bi×Bi. Six expression difference results were obtained by calculating the difference for each of the two group comparisons (diff1 vs 2; diff1 vs3; diff2 vs3; diff 2 vs 4; diff 3 vs4; and diff1 vs4). The cut off was adjusted to identify additional genes with differential expression patterns, by grouping genes into 4 expression patterns as follows:

Maternal genome driven-Brahman = $\text{diff23}/\text{diff14} > 0.8$, $\text{diff12}/\text{diff14} < 0.2$ and $\text{diff34}/\text{diff14} < 0.2$

Paternal genome driven-Angus = $\text{diff23}/\text{diff14} > 0.8$, $\text{diff13}/\text{diff14} < 0.2$ and $\text{diff24}/\text{diff14} < 0.2$

Angus dominant = $\text{diff23}/\text{diff14} < 0.2$, $\text{diff24}/\text{diff14} < 0.8$ and $\text{diff34}/\text{diff14} < 0.8$

Brahman dominant = $\text{diff23}/\text{diff14} < 0.2$, $\text{diff12}/\text{diff14} < 0.8$ and $\text{diff13}/\text{diff14} < 0.8$

For example, for a given gene to show either maternal or paternal genome driven expression, the average expression difference between two crossbred groups (2 vs 3) must be large and close to the average expression difference between two purebred group (1 vs 4). To ascertain whether maternal or paternal genome driven expression is driven by the breed type, we next checked whether the average expression of crossbred (Bi×Bt) is close to the expression of the Brahman purebred (Bi×Bi) or the Angus purebred (Bt×Bt). If the average expression of crossbred (Bi×Bt) was close to the expression of Bt×Bt, the expression difference between these two groups (denote as diff12) should be very small compared with the difference between two purebred groups (denote as diff14) and the difference in average expression of the crossbred (Bt×Bi) and average expression of Bi×Bi should also be small.

4.4 Results

4.4.1 Expression profile of five tissues

Samples were analysed from 5 tissues (brain, liver, lung, muscle and placenta) from the 4 cattle genotypes and from the 2 sexes, with 3 biological replicates for each, giving a total of 120 samples. On average, 60-100 million 100bp paired-end reads per sample passed quality control. Reads were aligned to the modified Brahman reference genome (UOA_brahman_1 plus non-PAR Y chromosome from UOA_angus_1) using hisat2 with default settings, giving an average mapping rate of 89%. The total number of expressed genes among samples ranged from 16,368 to 17,013 and showed no substantial variation between tissues. There was a high correlation coefficient between the pure breeds for each tissue (Supplementary figure 4.4a-e). There were 14,143 genes expressed in all tissues (Supplementary Figure 4.4f) with 5 genes consistently highly expressed in all five tissues: Insulin-Like Growth Factor 2 (*IGF2*), Eukaryotic Translation Elongation Factor 1 Alpha 1 (*EEF1A1*), Collagen Type III Alpha 1 Chain (*COL3A1*), Actin Beta (*ACTB*) and the paternally expressed gene (*PEG3*).

Multi-dimensional scaling analysis of the 5 tissues showed that each tissue forms a tight cluster, which is distinct and well separated from the others, irrespective of the genotype of origin (Figure 4.1a). After fitting a multi-factor model that accounted for tissue effects, the samples were separated by genetic groups in the first principle component (x-axis) and by sex in the second principle component (y-axis) (Figure 4.1b). The expression level of samples for each tissue showed the same pattern, with the 2 purebred groups clustering separately for all tissues, and the reciprocal crosses less clearly separated (Supplementary Figure 4.5a-e). The 20 most highly expressed genes in each tissue are reported in Supplementary Table 4.2.

4.4.2 Differential gene expression between purebred groups

There were 1,085, 1,495, 1,935, 2,515 and 2,645 DEGs between Angus and Brahman brain, placenta, lung, liver and muscle respectively. In each tissue, the number of up-regulated and down-regulated genes were similar. Notably, muscle had most DE genes

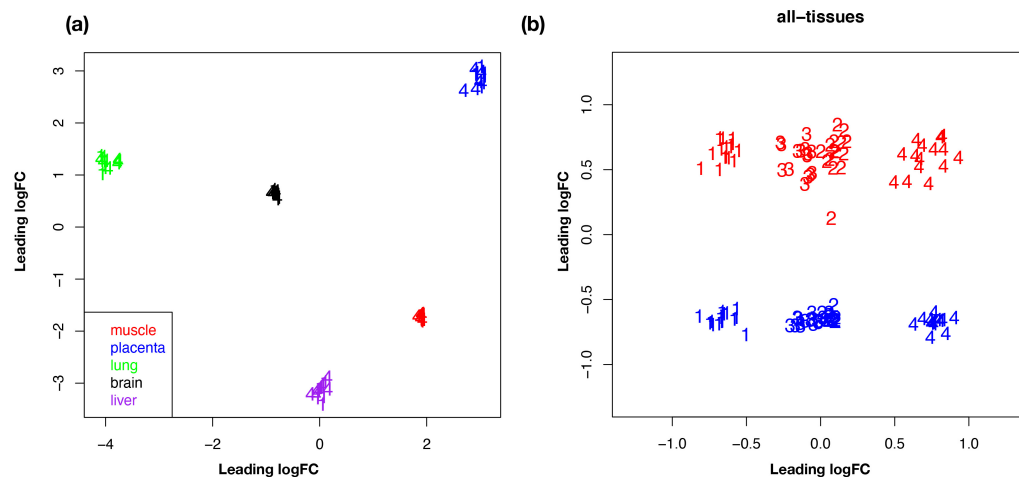


Figure 4.1: Multi-dimensional scaling (MDS) plot of sample expression profiles in five tissues.

a) The first two dimensions separate the samples by tissue type. b) After accounting for the tissue source, all samples are separated by genetic group in the first dimension (X-axis) and by sex in the second dimension (Y-axis). (1-pure Angus, 2-Angus X Brahman, 3-Brahman X Angus, 4-pure Brahman).

among tissues, but about 84% of DEGs showed a fold change (FC) < 2 in muscle while only ~62%-72% showed a FC < 2 in other tissues. The most significantly enriched gene ontology (GO) biological process and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in muscle included collagen metabolic process (GO:0032963); collagen fibril organization (GO:0030199); Amino sugar and nucleotide sugar metabolism (bta00520) and Glycine, serine and threonine metabolism (bta00260). Genes in all these pathways had higher expression in Angus than in Brahman.

4.4.3 DE genes common to all five tissues

There were 110 DE genes between Brahman and Angus common to all tissues, including 42 novel protein-coding genes and 18 lncRNAs (Figure 4.2a). The identity of the novel protein-coding genes was searched by aligning them to known genes in other cattle and ruminant reference genomes, which enabled 37 of the novel genes to be annotated, based on >90% sequence identity. Of the 87 annotated protein-coding genes that were DE between purebred animals in all five tissues, 83 were consistently up- or down-regulated with respect to genotype in all tissues. The 3 exceptions were Aldehyde Oxidase 1 (*AOX1*), Choline Dehydrogenase (*CHDH*), Syntaxin 11 (*STX11*), whose expression was

in a different direction (Angus vs Brahman) in the liver compared with the other 4 tissues.

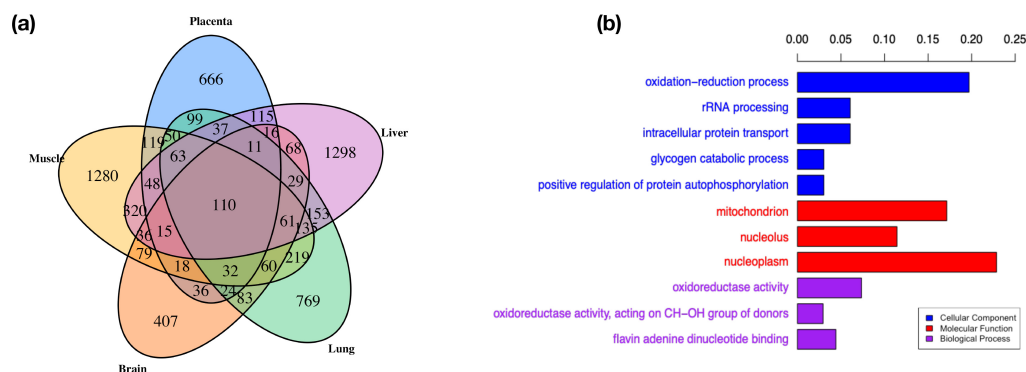


Figure 4.2: Venn diagram with numbers of differentially expressed genes across five tissues and their pathways

a) Venn diagram depicting the distribution of DE genes across five tissues at FDR cut off 0.05. b) Significantly enriched gene ontology terms for biological process (purple), Molecular function (red) and cellular component (blue) for 87 annotated common DE genes in all five tissues.

To obtain an insight into the fundamental biological relationships among genes that were differentially expressed between the two pure breeds in all five tissues, we performed GO and KEGG pathway analysis. The GO analysis showed that DEGs were significantly enriched in 10 GO terms, including oxidation-reduction process (GO:0055114), intracellular protein transport (GO:0006886), glycogen catabolic process (GO:0005980), positive regulation of protein autophosphorylation (GO:0031954) (Figure 4.2b).

4.4.4 Tissue-specific genes between purebred groups

There were 407, 666, 769, 1,298 and 1,280 tissue-specific DE genes between pure Angus and pure Brahman in brain, placenta, lung, liver and muscle, respectively with an FDR cut-off of <0.05 (Figure 4.2a). To select DE genes most strongly associated with each tissue, we further filtered the genes using absolute fold change (FC) ≥ 2 , leaving less than a third of the tissue-specific DE genes, i.e., 187, 328, 289, 388 and 191 DEGs in each of the tissues respectively. We performed GO biological process pathway enrichment analysis for these filtered DE genes for each tissue, which identified 54 GO terms for the tissue specific genes (Supplementary Table 4.3). The liver-specific DE genes were

enriched for 6 GO terms including ion binding and primary metabolic processes; both brain and muscle were enriched for 9 GO terms. Muscle was enriched for the collagen fibril organization pathway, while brain enriched pathways that included detection of stimulus and nervous system process. Lung was enriched for 10 GO terms, most of which were related to fundamental biological processes, including regulation of molecular function and cellular response to endogenous stimulus. Placenta-specific DE genes were linked to proton-transporting V-type ATPase and V1 domain small molecule metabolic process.

4.4.5 Differential gene expression between crossbred groups

When expression patterns between the reciprocal cross-bred groups were compared, only liver showed a substantial number of DEGs (2,473), while the other tissues had fewer than 20 DEGs each at $FDR < 0.05$. Among the 2,473 liver DE genes between crossbred-groups, only 143 DE had a fold change greater than 2. We performed GO biological process pathway enrichment analysis and KEGG pathway enrichment analysis for the protein coding DE genes with large fold change. The GO analysis showed that DEGs were significantly enriched in 6 GO terms, including: macromolecule metabolic process (GO:0043170), primary metabolic process (GO:0044238), cellular metabolic process (GO:0044237), metabolic process (GO:0008152), nitrogen compound metabolic process (GO:0006807) and organic substance metabolic process (GO:0071704) which are all involved in metabolic processes. The only significantly enriched KEGG pathway was Metabolic pathways (path: bta01100).

To explore similarities in expression patterns between the reciprocal crosses and the pure breed individuals we carried out pairwise comparisons of the DE genes in liver for the 4 genetic groups. The number of DEGs detected comparing Bt×Bi vs Bi×Bi was 1,276 which is five times greater than the number of DEGs from the Bt×Bi vs Bt×Bt comparison (219). The number of DEGs detected for Bi×Bt vs Bt×Bt was 317 which is twice the number of DEGs from Bi×Bt vs Bi×Bi (150).

4.4.6 Expression pattern of DEGs from the purebred cattle in comparison with crossbred groups

The expression pattern of genes which were DEG in the purebred animals were studied in the reciprocal crossbred groups. The majority (~90%) of these genes showed an additive expression pattern where both paternal and maternal genomes contributed to the gene expression levels in the crossbred groups (Figure 4.3a). To test whether there were parent-of-origin effects, the average expression of each gene in the 4 groups was calculated: pure Angus group (Bt×Bt), Brahman/Angus group (Bi×Bt), Angus/Brahman group (Bt×Bi) and pure Brahman group (Bi×Bi).

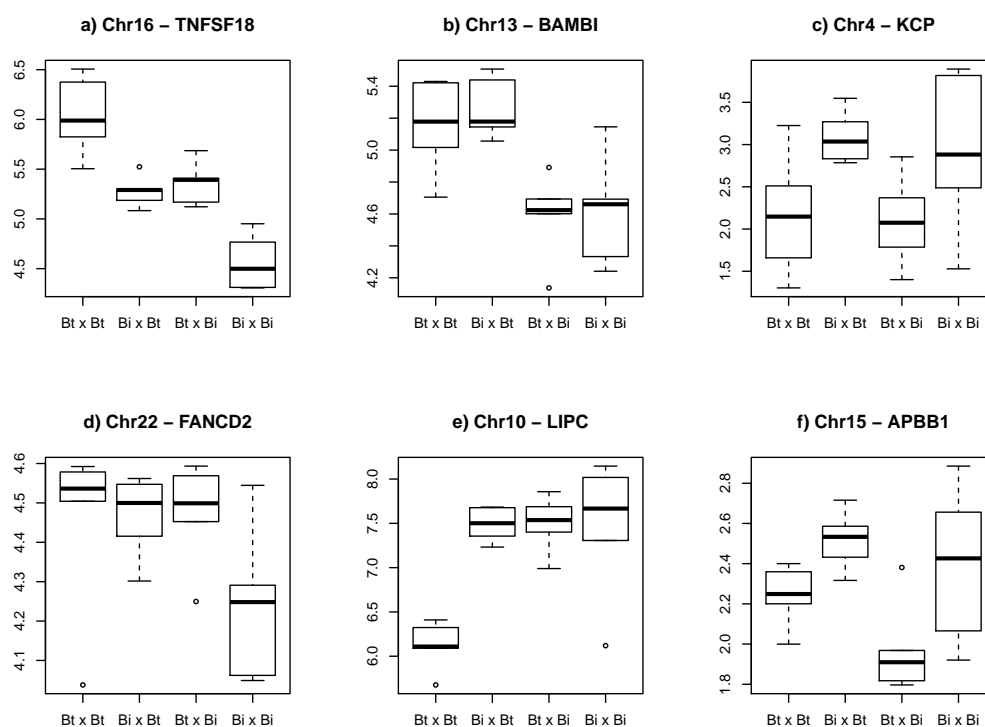


Figure 4.3: Examples of expression patterns among genotype groups.

Boxplots illustrate the different expression patterns observed among the 4 genetics groups: Bt×Bt, Bi×Bt, Bt×Bi and Bi×Bi. Y-axis is expression level (counts per million) in a log₂ scale. a) Taurus driven additive expression, irrespective of parent. b) Maternal genome driven Taurine dominance. c) Paternal genome driven, indicine dominance. d) Taurine dominant. e) Indicine dominant. f) complex inheritance.

A ratio of average gene expression between the 4 groups (pairwise comparison) was obtained and ~5% of DEGs in each tissue showed a parent of origin effect (Table 4.1): Angus or Brahman dominant and maternal/paternal driven expression (Figure 4.3b-e).

Table 4.1: Number of genes showing a parent of origin effect on expression patterns in five tissues.

	Maternal genome driven - Brahman	Paternal genome driven - Angus	Dominant Angus	Dominant Brahman
Brain	0	14	34	15
Liver	23	89	95	60
Lung	6	29	64	20
Muscle	27	43	43	30
Placenta	5	23	37	15

As only liver had a substantial number of DEGs in the crossbred comparison, we examined the overlap of DEGs between the crossbred groups with DEGs between purebred groups. Interestingly, only 310 DEGs overlapped. Expression patterns of these common genes had either maternal or paternal genome driven patterns. GO analysis showed that the overlapping DEGs were significantly enriched in 19 GO terms including positive regulation of cellular metabolic process, positive regulation of nitrogen compound metabolic process and membrane-enclosed lumen. We further examined the expression direction of the DEGs involved in these significantly pathways, which showed they had higher expression in the purebred Angus (Bt) compared with the Brahman (Bi). The DEGs between crossbred groups, fall into three general categories of co-dominant, dominant and recessive expression patterns, with dominance in some cases driven by either the male or female inherited allele (Figure 4.3).

4.5 Discussion

The study of gene expression in prenatal development will help us to understand the regulation of fetal tissue-specific growth and development. In this study we observed substantial differences in expression between breeds of cattle from the two genetically distinct sub-species *Bos taurus taurus* and *Bos taurus indicus*. In addition, we observed differential expression of genes in reciprocal crosses between these subspecies, some of which revealed parent-of-origin effects in gene expression in five tissues at mid-gestation.

Five genes that had high levels of expression in all five tissues (*IGF2*, *EEF1A1*, *COL3A1*, *ACTB* and *PEG3*), are likely to play important roles at mid-gestation. All five genes

play a crucial role in embryonic development and fetal growth, the consequences of loss-of-function mutations in these genes cause developmental delay and several diseases including intellectual disability, immune system abnormalities, cerebral abnormalities and abnormally large abdominal organs (Curley et al., 2004; Azzi et al., 2014; Abbas et al., 2015; Cuvertino et al., 2017; Horn et al., 2017). *EEF1A1* is a member of the eukaryotic elongation factor family that regulates protein synthesis, and has been shown to be expressed in brain, placenta, lung, liver, kidney, and pancreas in human adults (Hamey and Wilkins, 2018). *COL3A1* is expressed in extensible connective tissues, such as skin and lung, mutation of *COL3A1* has been linked to vascular type disease (Cortini et al., 2017).

Genomic imprinting has been described in various mammalian species and results in a biased level of expression of one of the 2 gene copies, depending on the parent of origin. Both insulin-like growth factor (*IGF2*) and paternally expressed gene 3 (*PEG3*) are imprinted genes, which have been found to be paternally expressed during prenatal life then expression declines rapidly after birth (Bergman et al., 2013). Both these genes have been shown to play an important role in controlling fetal growth rate and nurturing behaviours in mammals. In the present study, *IGF2* and *PEG3* were highly expressed in all samples across the 4 pure and crossbred groups in all five tissues, suggesting that both *PEG3* and *IGF2* play an important role at mid-gestation. However, there was no evidence of parent of origin effects on the overall level of expression of *PEG3* and *IGF2*, as expression did not differ between breeds or the direction of the cross. We were unable to assign transcripts to a parent of origin to test for imprinting, because with short read sequences too few could be assigned to the parental chromosome. Additional data such as single cell RNA-seq and Iso-seq would allow the allelic expression to be better tested.

Other highly expressed genes showed tissue-specific expression patterns which were related to tissue function. Alpha-Fetoprotein (*AFP*) had liver-specific expression and encodes a major plasma protein produced by the liver during fetal development (Petit et al., 2009). Two genes that were highly expressed in the muscle were the muscle structural protein genes Myosin Heavy Chain 3 (*MYH3*) and Myosin Binding Protein C, Slow Type

(*MYBPC1*) (Ha et al., 2013; Zieba et al., 2017). Genes that play an important role in neurodevelopment including Adenylate Cyclase 1 (*ADCY1*), Stathmin 2 (*STMN2*) and Tubulin Beta 3 Class III (*TUBB3*) were highly expressed and specific to the brain (Wang et al., 2004; Fukumura et al., 2016; Wang et al., 2019). All of these genes showed a similar, high level of expression in both pure breeds and their crosses. The lung was the only tissue that did not have any highly expressed tissue-specific genes at this developmental stage.

Intrauterine stress during fetal development increases the risk of adult disease. Increased oxidative stress during embryonic and fetal growth may be caused by multiple conditions (Thompson and Al-Hasan, 2012) and may affect transcription factors which can change the expression of key genes at developmental stage (Dennery, 2004). From the GO pathway analysis in the current study, the oxidation-reduction process and oxidoreductase activity were found to be significantly associated with the DEGs between the two pure breeds that were in common to all five tissues.

Heat shock leads to oxidative stress and reduced production performance, which has been studied in *Bos taurus indicus* (Fedyaeva et al., 2014). Oxidative damage to cells and mitochondria has been shown to be caused by the changes in the steady-state concentration of free radicals during heat stress (Belhadj Slimen et al., 2016). A study on the effects of oxidative stress on cattle fertility suggest *Bos taurus taurus* bulls have a higher level of reactive oxygen species (ROS) in their semen than *Bos taurus indicus* bulls in tropical areas (Nichi et al., 2006). These high level of ROS have been suggested as a cause of major sperm defects in the *Bos taurus taurus* bulls (Nichi et al., 2006). In our study, *TXNRD2*, a mitochondrial protein that scavenges reactive oxygen species had a higher level of expression in Brahman than Angus in all tissues. Suggesting that *TXNRD2* mediated protection of mitochondrial function may help indicine cattle adapt to the hot environment.

HSD11B1L is a protein which catalyses the interconversion of inactive to active glucocorticoids, e.g., the conversion of inactive cortisone to the active forms corticosterone, cortisol, which are key hormones that regulate a variety of physiologic responses to stress through the hypothalamus-pituitary-adrenal (HPA) axis that is responsible for the adaptation of stress responses to restore homeostasis (Walker et al., 2015). *HSD11B1L* had higher lev-

els of expression in all Brahman tissues, which may allow indicus cattle to respond more rapidly to stress.

Most of the genes that were DE in all five tissues showed changes in the level of expression in the same direction in Angus and Brahman for all tissues. *AOXI* and *CHDH* were the exceptions, which had a different direction of expression in the liver compared with the other 4 tissues. The liver plays an important role in metabolic processes and in immune system function which affects the response to many diseases (Chang et al., 2017; MacParland et al., 2018). Expression of *AOXI* produces hydrogen peroxide and catalyses the formation of superoxide. We found that the expression of *AOXI* was high in all Angus tissues except in liver where there was a higher level in Brahman compared with Angus. Levels of *AOXI* have been found to increase in mouse liver in relation to increased immune response following infection (Maeda et al., 2012) suggesting a role in immune response by stimulating host immunity, inflammation and coagulation. Indicine cattle are generally less susceptible to disease than taurine cattle (Mackinnon et al., 1991; Vajana et al., 2018). For example, indicine cattle are more resistant to ticks (Franzin et al., 2017) and tuberculosis (Vordermeier et al., 2012). Interestingly *AOXI* had lower levels of expression in Brahman than Angus in tissues other than liver. The significance of this is not known.

The GO terms including genes that were DE between purebreds in this study showed that those involved in metabolic processes generally had significantly higher expression in Angus compared to Brahman. Low metabolic rate is associated with thermotolerance of *Bos taurus indicus* (Hansen, 2004). Interestingly, the genes that were DE between the liver of the purebred fetus that were also differentially expressed between the reciprocal crossbred fetus showed a higher expression when the sire was taurine for both sexes. For example, a critical nuclear receptor *NR4A1* had a higher level of expression in pure Angus and also in the crossbred fetus when the sire was Angus. *NR4A1* is involved in inflammation, apoptosis, and glucose metabolism and also regulates a paternal imprinted SNRPN, which affects neurological and spine development (Li et al., 2016). *NR4A1* regulates energetic competence of mitochondria and promotes neuronal plasticity. However, studies in animal models and of neuropathologies in humans have shown that sustained expression results in

individuals being sensitive to chronic stress (Jeanneteau et al., 2018). Therefore, the higher levels of expression in Angus, may mean that they are less able to cope with stress, such as heat and drought conditions than the Brahman.

Up to now the development of indicine and taurine composites has not taken into account which of the types is use as sire or dam. Our data suggests the paternal genome has an effect on the expression of genes involved in e.g., metabolic processes, stress response and neuronal development, which should be taken into account.

In conclusion, this study identified a large number of genes that showed a significant breed difference in expression in each tissue. These genes were found to participate in pathways related to tissue-specific function. Genes that were differentially expressed between Angus and Brahman in all tissues were found to relate to functions such as immune response and stress response. This study also identified genes that putatively have parent or breed of origin-controlled expression patterns. Exploring these further would require long read Iso-seq data with parentally phased whole genome sequence genotypes to resolve haplotype specific expression. The data provide a basis for future research on parental genome effects underlying phenotypic differences in cattle fetal development. Taking these factors into account may improve the welfare and productivity of crossbred cattle in tropical environments.

References

- Abbas, W., Kumar, A., and Herbein, G., 2015. The eef1a proteins: at the crossroads of oncogenesis, apoptosis, and viral infections. *Frontiers in oncology*, 5, p.75.
- Alexandre, P.A., Kogelman, L.J., Santana, M.H., Passarelli, D., Pulz, L.H., Fantinato-Neto, P., Silva, P.L., Leme, P.R., Strefezzi, R.F., and Coutinho, L.L., 2015. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. *Bmc genomics*, 16(1), p.1073.
- Andrew, S., 2010. *Fastqc, a quality control tool for high throughput sequence data*. [Online]. Generic.
- Azzi, S., Habib, W.A., and Netchine, I., 2014. Beckwith–wiedemann and russell–silver syndromes: from new molecular insights to the comprehension of imprinting regulation. *Current opinion in endocrinology, diabetes and obesity*, 21(1), pp.30–38.
- Belhadj Slimen, I., Najar, T., Ghram, A., and Abdrrabba, M., 2016. Heat stress effects on livestock: molecular, cellular and metabolic aspects, a review. *Journal of animal physiology and animal nutrition*, 100(3), pp.401–412.
- Bergman, D., Halje, M., Nordin, M., and Engström, W., 2013. Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology*, 59(3), pp.240–249.
- Berton, M.P., Fonseca, L.F., Gimenez, D.F., Utembergue, B.L., Cesar, A.S., Coutinho, L.L., Lemos, M.V.A. de, Aboujaoude, C., Pereira, A.S., and Rafael, M.d.O., 2016. Gene expression profile of intramuscular muscle in nellore cattle with extreme values of fatty acid. *Bmc genomics*, 17(1), p.972.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L., 2009. Blast+: architecture and applications. *Bmc bioinformatics*, 10(1), p.421.

- Chang, H., Meng, H.-y., Liu, S.-m., Wang, Y., Yang, X.-x., Lu, F., and Wang, H.-y., 2017. Identification of key metabolic changes during liver fibrosis progression in rats using a urine and serum metabolomics approach. *Scientific reports*, 7(1), pp.1–12.
- Consortium, B.H. et al., 2009. The genetic history of cattle. *Science*, 324, pp.528–532.
- Cortini, F., Marinelli, B., Romi, S., Seresini, A., Pesatori, A.C., Seia, M., Montano, N., and Bassotti, A., 2017. A new col3a1 mutation in ehlers-danlos syndrome vascular type with different phenotypes in the same family. *Vascular and endovascular surgery*, 51(3), pp.141–145.
- Cui, X., Hou, Y., Yang, S., Xie, Y., Zhang, S., Zhang, Y., Zhang, Q., Lu, X., Liu, G.E., and Sun, D., 2014. Transcriptional profiling of mammary gland in holstein cows with extremely different milk protein and fat percentage using rna sequencing. *Bmc genomics*, 15(1), p.226.
- Curley, J.P., Barton, S., Surani, A., and Keverne, E.B., 2004. Coadaptation in mother and infant regulated by a paternally expressed imprinted gene. *Proceedings of the royal society of london. series b: biological sciences*, 271(1545), pp.1303–1309.
- Cuvertino, S., Stuart, H.M., Chandler, K.E., Roberts, N.A., Armstrong, R., Bernardini, L., Bhaskar, S., Callewaert, B., Clayton-Smith, J., and Davalillo, C.H., 2017. Actb loss-of-function mutations result in a pleiotropic developmental disorder. *The american journal of human genetics*, 101(6), pp.1021–1033.
- Dennerly, P.A., 2004. Role of redox in fetal development and neonatal diseases. *Antioxidants and redox signaling*, 6(1), pp.147–153.
- Durinck, S., Spellman, P.T., Birney, E., and Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8), p.1184.
- Fedyaeva, A., Stepanov, A., Lyubushkina, I., Pobezhimova, T., and Rikhvanov, E., 2014. Heat shock induces production of reactive oxygen species and increases inner mito-

- chondrial membrane potential in winter wheat cells. *Biochemistry (moscow)*, 79(11), pp.1202–1210.
- Franzin, A.M., Maruyama, S.R., Garcia, G.R., Oliveira, R.P., Ribeiro, J.M.C., Bishop, R., Maia, A.A.M., Moré, D.D., Ferreira, B.R., and Miranda Santos, I.K.F. de, 2017. Immune and biochemical responses in skin differ between bovine hosts genetically susceptible and resistant to the cattle tick *Rhipicephalus microplus*. *Parasites & vectors*, 10(1), p.51.
- Fukumura, S., Kato, M., Kawamura, K., Tsuzuki, A., and Tsutsumi, H., 2016. A mutation in the tubulin-encoding *tubb3* gene causes complex cortical malformations and unilateral hypohidrosis. *Child neurology open*, 3, p.2329048X16665758.
- Ha, K., Buchan, J.G., Alvarado, D.M., McCall, K., Vydyanath, A., Luther, P.K., Goldsmith, M.I., Dobbs, M.B., and Gurnett, C.A., 2013. *Mybpc1* mutations impair skeletal muscle function in zebrafish models of arthrogyrosis. *Human molecular genetics*, 22(24), pp.4967–4977.
- Hamey, J.J. and Wilkins, M.R., 2018. Methylation of elongation factor 1a: where, who, and why? *Trends in biochemical sciences*, 43(3), pp.211–223.
- Hansen, P., 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Animal reproduction science*, 82, pp.349–360.
- Hayes, B.J. and Daetwyler, H.D., 2019. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences* [Online], 7(1). PMID: 30508490, pp.89–102. eprint: <https://doi.org/10.1146/annurev-animal-020518-115024>.
- Horn, D., Siebert, E., Seidel, U., Rost, I., Mayer, K., Abou Jamra, R., Mitter, D., and Kornak, U., 2017. Biallelic *col3a1* mutations result in a clinical spectrum of specific structural brain anomalies and connective tissue abnormalities. *American journal of medical genetics part a*, 173(9), pp.2534–2538.

- Jeanneteau, F., Barrère, C., Vos, M., De Vries, C.J., Rouillard, C., Levesque, D., Dromard, Y., Moisan, M.-P., Duric, V., and Franklin, T.C., 2018. The stress-induced transcription factor nr4a1 adjusts mitochondrial function and synapse number in prefrontal cortex. *Journal of neuroscience*, 38(6), pp.1335–1350.
- Kim, D., Langmead, B., and Salzberg, S., 2015. Hisat2. *Nature methods*.
- Krog, C.H., Agerholm, J.S., and Nielsen, S.S., 2018. Fetal age assessment for holstein cattle. *Plos one*, 13(11).
- Law, C.W., Chen, Y., Shi, W., and Smyth, G.K., 2014. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2), R29.
- Li, H., Zhao, P., Xu, Q., Shan, S., Hu, C., Qiu, Z., and Xu, X., 2016. The autism-related gene snrpn regulates cortical and spine development via controlling nuclear receptor nr4a1. *Scientific reports*, 6, p.29878.
- Liao, Y., Smyth, G.K., and Shi, W., 2019. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. *Nucleic acids research*, 47(8), e47–e47.
- Liu, R., Holik, A.Z., Su, S., Jansz, N., Chen, K., Leong, H.S., Blewitt, M.E., Asselin-Labat, M.-L., Smyth, G.K., and Ritchie, M.E., 2015. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic acids research*, 43(15), e97–e97.
- Mackinnon, M., Meyer, K., and Hetzel, D., 1991. Genetic variation and covariation for growth, parasite resistance and heat tolerance in tropical cattle. *Livestock production science*, 27(2-3), pp.105–122.
- MacParland, S.A., Liu, J.C., Ma, X.-Z., Innes, B.T., Bartczak, A.M., Gage, B.K., Manuel, J., Khuu, N., Echeverri, J., and Linares, I., 2018. Single cell rna sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature communications*, 9(1), pp.1–21.

- Maeda, K., Ohno, T., Igarashi, S., Yoshimura, T., Yamashiro, K., and Sakai, M., 2012. Aldehyde oxidase 1 gene is regulated by nrf2 pathway. *Gene*, 505(2), pp.374–378.
- Mao, W., Albrecht, E., Teuscher, F., Yang, Q., Zhao, R., and Wegner, J., 2008. Growth- and breed-related changes of fetal development in cattle. *Asian-australasian journal of animal sciences*, 21(5), pp.640–647.
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet. journal*, 17(1), pp.10–12.
- Menéndez-Buxadera, A., Palacios-Espinosa, A., Espinosa-Villavicencio, J.L., and Guerra-Iglesias, D., 2016. Genotype environment interactions for milk production traits in holstein and crossbred holstein-zebu cattle populations estimated by a character state multibreed model. *Livestock science*, 191, pp.108–116.
- Mukiibi, R., Vinsky, M., Keogh, K.A., Fitzsimmons, C., Stothard, P., Waters, S.M., and Li, C., 2018. Transcriptome analyses reveal reduced hepatic lipid synthesis and accumulation in more feed efficient beef cattle. *Scientific reports*, 8(1), pp.1–12.
- Nichi, M., Bols, P., Züge, R.M., Barnabe, V.H., Goovaerts, I., Barnabe, R.C., and Cortada, C.N.M., 2006. Seasonal variation in semen quality in bos indicus and bos taurus bulls raised under tropical conditions. *Theriogenology*, 66(4), pp.822–828.
- Petit, F.M., Hébert, M., Picone, O., Brisset, S., Maurin, M.-L., Parisot, F., Capel, L., Benattar, C., Sénat, M.-V., and Tachdjian, G., 2009. A new mutation in the afp gene responsible for a total absence of alpha feto-protein on second trimester maternal serum screening for down syndrome. *European journal of human genetics*, 17(3), pp.387–390.
- Pitt, D., Sevane, N., Nicolazzi, E.L., MacHugh, D.E., Park, S.D.E., Colli, L., Martinez, R., Bruford, M.W., and Orozco-terWengel, P., 2019. Domestication of cattle: two or three events? *Evol appl* [Online], 12(1), pp.123–136.
- Porto-Neto, L.R., Reverter, A., Prayaga, K.C., Chan, E.K., Johnston, D.J., Hawken, R.J., Fordyce, G., Garcia, J.F., Sonstegard, T.S., and Bolormaa, S., 2014. The genetic architecture of climatic adaptation of tropical cattle. *Plos one*, 9(11).

- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K., 2015. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7), e47–e47.
- Robinson, M.D. and Oshlack, A., 2010. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3), R25.
- Sheng, X., Ni, H., Liu, Y., Li, J., Zhang, L., and Guo, Y., 2014. Rna-seq analysis of bovine intramuscular, subcutaneous and perirenal adipose tissues. *Molecular biology reports*, 41(3), pp.1631–1637.
- Thompson, L.P. and Al-Hasan, Y., 2012. Impact of oxidative stress in fetal programming. *Journal of pregnancy*, 2012.
- Vajana, E., Barbato, M., Colli, L., Milanese, M., Rochat, E., Fabrizi, E., Mukasa, C., Del Corvo, M., Masembe, C., and Muwanika, V.B., 2018. Combining landscape genomics and ecological modelling to investigate local adaptation of indigenous ugandan cattle to east coast fever. *Frontiers in genetics*, 9, p.385.
- Vordermeier, M., Ameni, G., Berg, S., Bishop, R., Robertson, B.D., Aseffa, A., Hewinson, R.G., and Young, D.B., 2012. The influence of cattle breed on susceptibility to bovine tuberculosis in ethiopia. *Comparative immunology, microbiology and infectious diseases*, 35(3), pp.227–232.
- Walker, J.J., Spiga, F., Gupta, R., Zhao, Z., Lightman, S., and Terry, J., 2015. Rapid intra-adrenal feedback regulation of glucocorticoid synthesis. *Journal of the royal society interface*, 12(102), p.20140875.
- Wang, H., Ferguson, G.D., Pineda, V.V., Cundiff, P.E., and Storm, D.R., 2004. Overexpression of type-1 adenylyl cyclase in mouse forebrain enhances recognition memory and ltp. *Nature neuroscience*, 7(6), pp.635–642.
- Wang, Q., Zhang, Y., Wang, M., Song, W.-M., Shen, Q., McKenzie, A., Choi, I., Zhou, X., Pan, P.-Y., and Yue, Z., 2019. The landscape of multiscale transcriptomic networks and key regulators in parkinson's disease. *Nature communications*, 10(1), pp.1–15.

- Xiang, R., Ghanipoor-Samami, M., Johns, W.H., Eindorf, T., Rutley, D.L., Kruk, Z.A., Fitzsimmons, C.J., Thomsen, D.A., Roberts, C.T., and Burns, B.M., 2013. Maternal and paternal genomes differentially affect myofibre characteristics and muscle weights of bovine fetuses at midgestation. *Plos one*, 8(1).
- Xiang, R., Lee, A.M., Eindorf, T., Javadmanesh, A., Ghanipoor-Samami, M., Gugger, M., Fitzsimmons, C.J., Kruk, Z.A., Pitchford, W.S., and Leviton, A.J., 2014. Widespread differential maternal and paternal genome effects on fetal bone phenotype at mid-gestation. *Journal of bone and mineral research*, 29(11), pp.2392–2404.
- Zeng, L., Cao, Y., Wu, Z., Huang, M., Zhang, G., Lei, C., and Zhao, Y., 2019. A missense mutation of the *hspb7* gene associated with heat tolerance in chinese indicine cattle. *Animals (basel)* [Online], 9(8).
- Zieba, J., Zhang, W., Chong, J.X., Forlenza, K.N., Martin, J.H., Heard, K., Grange, D.K., Butler, M.G., Kleefstra, T., and Lachman, R.S., 2017. A postnatal role for embryonic myosin revealed by *myh3* mutations that alter $\text{tg}\beta$ signaling and cause autosomal dominant spondylacropotarsal synostosis. *Scientific reports*, 7, p.41803.

4.6 Supplementary Figures

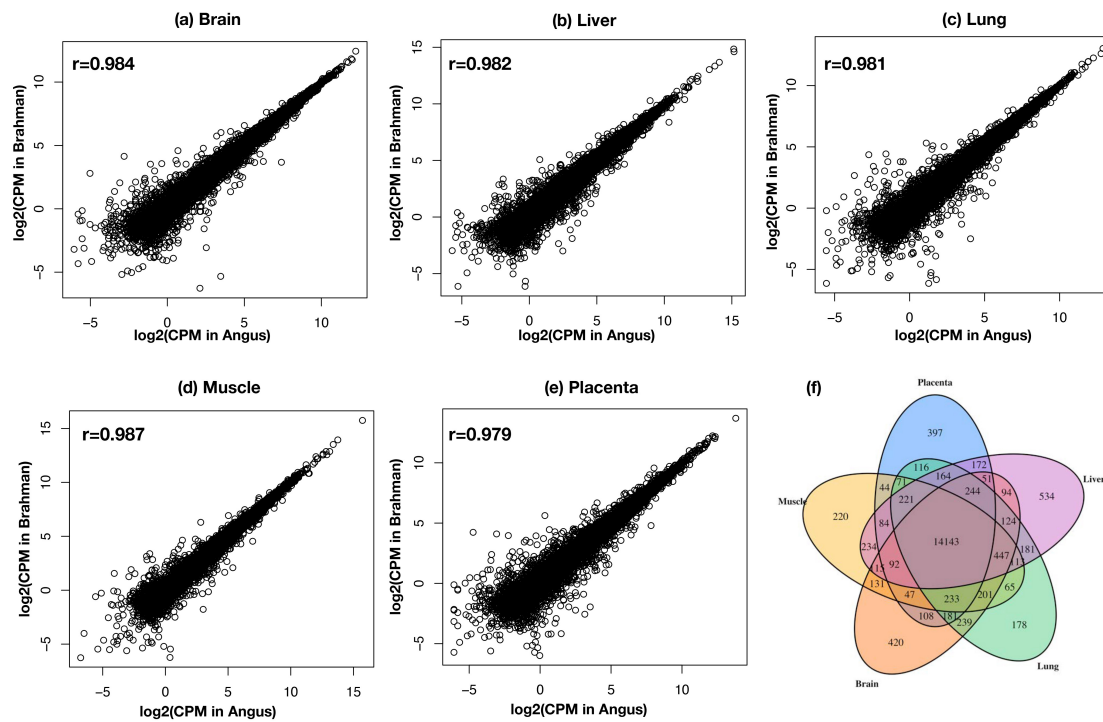


Figure 4.4: Comparison of gene expression levels in five tissues for pure Angus and Brahman.

(a)-(e) The X and Y axes plot the gene expression counts (log2 count per million) in Angus vs Brahman in Brain, Liver, Lung, Muscle and Placenta, respectively. f) Venn diagram shows the overlap of expressed genes in five tissues.

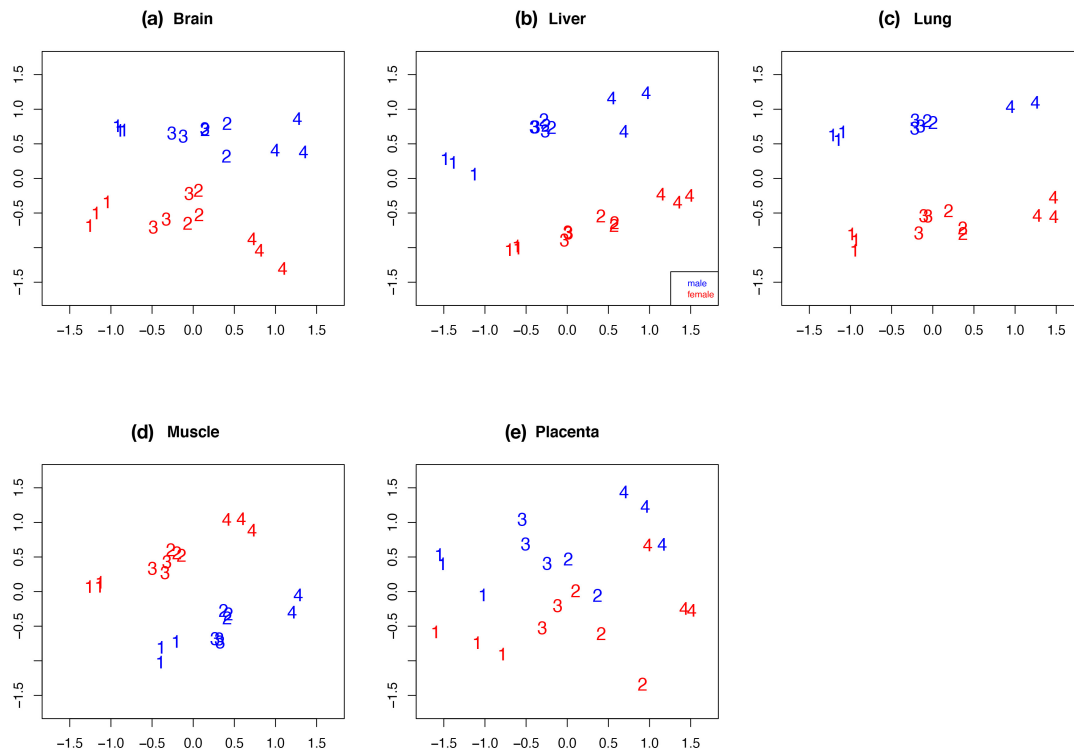


Figure 4.5: Multi-dimensional scaling plots reveals genetic group difference in gene expression profiles in each tissue.

Male samples are in blue and female samples are in red. The X and Y axes are in log₂ fold changes. a) Brain; b) Liver; c) Lung; d) Muscle; e) Placenta.

4.7 Supplementary Tables

Table 4.2: Highly expressed genes (average CPM) in five tissues.

Chr	Start	End	Symbol	Placenta	Muscle	Brain	Lung	Liver
1	80335480	80342557	AHSG	14.929	-3.754	NA	NA	NA
6	88039295	88057657	ALB	14.9	0.486	-0.909	1.919	-2.326
21	58943373	58955269	SERPINA1	13.913	-3.607	NA	NA	-1.906
6	88065424	88087062	AFP	13.441	-2.142	NA	NA	NA
1	134958638	135044784	LOC113895043	13.238	4.86	5.534	5.571	5.627
15	35604141	35605767	LOC113905582	12.603	5.23	8.305	5.806	7.039
17	69741253	69749107	FGB	12.258	-0.229	NA	NA	NA
X	128075530	128539068	GPC3	11.976	9.617	8.209	3.616	3.519
11	77137851	77179834	APOB	11.706	-3.665	NA	NA	NA
13	67579296	67611649	ITIH2	11.619	-0.225	-0.728	-0.015	3.293
4	107686367	107699081	PEG10	11.593	10.4	8.942	10.98	9.401
6	86528024	86580214	GC	11.546	-2.592	NA	NA	NA
17	69774759	69782875	FGA	11.347	-4.928	NA	NA	NA
18	1704731	1728155	PEG3	11.106	7.046	6.564	8.485	12.149
1	80284653	80299937	FETUB	10.979	-3.631	NA	NA	NA
17	69794842	69802922	FGG	10.934	-4.045	NA	NA	1.181
22	12700318	12714256	ITIH3	10.749	-0.63	-0.663	-1.782	NA
3	8319332	8320657	APOA2	10.721	-4.534	0.658	NA	NA
8	103307653	103320100	AMBP	10.693	-3.331	NA	NA	NA
2	63594578	64814251	NCKAP5	10.653	10.528	5.311	5.598	2.816
2	17900389	18176615	TTN	4.634	15.727	3.013	2.785	2.866
X	30451046	33107990	DMD	8.267	13.441	8.381	6.846	2.877
19	30444588	30474852	LOC113877399	-1.446	12.993	-1.684	1.411	NA
2	7134607	7174433	COL3A1	9.676	12.754	12.758	6.555	10.215
2	44319518	44538758	NEB	3.667	12.548	2.961	4.688	5.465
29	38930699	41249691	DLG2	5.372	12.196	1.791	9.045	2.102
19	30643315	30664308	MYH3	0.532	12.154	3.09	1.088	-0.782
5	54748044	54849459	MYBPC1	-0.051	12.141	-1.363	-0.035	NA
21	65568371	65603630	LOC113879939	10.395	11.705	12.26	12.314	10.88
6	50128676	50604276	PCDH7	7.28	11.41	6.452	7.343	2.046
10	42742821	43553521	RORA	8.837	11.308	6.314	5.461	1.893
X	92746655	92875008	NRK	1.224	11.204	3.651	2.167	9.634
28	22239873	24150230	CTNNA3	5.962	11.201	1.763	3.417	1.248
19	37532052	37550039	COL1A1	6.889	11.175	11.645	6.863	9.511
25	11541914	12757716	AUTS2	10.597	11.135	6.311	7.035	2.784
12	23965480	24000775	POSTN	8.166	11.133	6.004	6.751	5.822
22	56132397	56935219	RBMS3	9.253	10.947	6.311	3.87	3.695
29	1746722	1774493	IGF2	10.482	10.878	13.205	4.34	11.664
8	29599753	29857371	NFIB	8.8	10.829	7.936	7.456	3.901
3	84324798	84725574	NFIA	10.208	10.779	6.504	6.101	3.859

Continued on next page

Table 4.2 – continued from previous page

Chr	Start	End	Symbol	Placenta	Muscle	Brain	Lung	Liver
7	22236998	22416166	MEF2C	6.866	10.746	5.511	10.248	4.128
21	43272612	44230078	NPAS3	6.363	10.744	4.187	5.651	2.017
10	46875925	47269137	TCF12	9.578	10.662	7.915	6.745	5.836
29	10206701	10299311	AHNAK	8.278	10.611	10.745	6.121	10.72
4	61981452	62937507	IMMP2L	10.062	10.566	3.949	4.453	4.469
24	54733068	55118668	TCF4	8.247	10.562	7.75	8.818	4.698
9	13105107	13111604	EEF1A1	10.178	10.481	11.934	10.227	10.209
13	25895716	25963127	LOC113903214	9.212	9.695	11.886	10.549	10.001
3	116885591	116976466	COL6A3	6.618	8.849	11.643	4.844	6.99
3	106719811	107060340	MACF1	9.685	10.019	11.414	9.696	9.012
14	77275253	77285338	CA3	2.685	8.694	11.064	5.286	-2.75
2	103263988	103333284	FN1	9.941	9.555	11.013	7.265	8.74
7	90600521	90609410	EEF2	8.802	9.151	11.011	10.726	9.881
11	36862372	37073566	SPTBN1	9.265	10.025	10.885	10.559	8.071
23	18804260	18809940	HSP90AB1	7.997	8.666	10.866	10.514	9.81
12	84456017	84589330	COL4A1	7.473	9.103	10.811	8.583	9.888
22	48136990	48150444	RPSA	8.963	8.509	10.792	9.674	9.475
25	40234488	40252770	SRRM2	8.099	7.255	10.731	9.463	9.974
7	47906558	47929571	SPARC	7.816	10.226	10.722	8.584	9.775
25	41018318	41020575	RPS2	8.849	8.502	10.719	10.326	9.769
10	7100359	7105068	RPL4	8.529	8.864	10.691	9.43	9.543
5	89638440	89642890	TUBA1A	4.612	6.865	7.452	11.9	5.499
10	101904614	101915069	CALM1	6.627	6.746	8.542	11.861	9.138
14	43120345	43178570	STMN2	2.603	NA	-0.228	11.634	NA
20	9245810	9339037	MAP1B	4.617	7.897	6.446	11.627	3.332
19	52342792	52345664	ACTG1	7.272	7.585	10.64	11.313	11.584
2	126790948	126821568	STMN1	5.546	6.155	8.106	11.272	5.989
11	98240094	98298529	SPTAN1	7.28	7.573	9.547	11.123	8.743
8	73721987	73840519	DPYSL2	6.129	6.892	8.201	11.063	5.419
7	52092400	52209325	DPYSL3	4.803	7.97	7.532	11.06	4.687
23	29182886	29187332	TUBB	8.038	7.635	9.789	10.943	8.969
2	97178044	97475943	MAP2	7.263	5.577	6.924	10.936	2.474
25	8149068	8173718	YWHAG	6.027	7.13	7.421	10.905	9.131
18	51083202	51091539	TUBB3	NA	-1.662	1.154	10.878	0.171
19	46939392	47063344	MAPT	2.146	3.266	6.263	10.823	1.407
4	43464704	43571682	ADCY1	2.465	1.776	-0.52	10.74	NA
13	29265820	29273580	EEF1A2	NA	2.688	NA	10.696	NA
16	78765060	78809629	KIF21B	2.991	1.867	6.916	10.637	3.006
29	13504876	13514381	LOC113886273	NA	-0.708	NA	-2.468	13.7
29	12504822	12514216	LOC113885740	NA	-3.081	NA	-3.506	12.242
23	36300765	36313454	LOC113881926	NA	NA	NA	-3.39	12.17
15	61655547	61667673	LOC113905846	NA	NA	NA	-3.833	12.143

Continued on next page

Table 4.2 – continued from previous page

Chr	Start	End	Symbol	Placenta	Muscle	Brain	Lung	Liver
13	9247781	9254044	LOC113902973	0.234	NA	NA	-3.814	11.955
23	51784919	51794891	LOC113881729	2.097	1.19	3.738	2.001	11.762
X	84976087	85002158	CAPN6	3.377	8.347	7.452	0.206	11.702
25	3736097	3739524	ACTB	9.192	8.432	10.484	10.126	11.67
6	84881046	84938250	SULT1E1	7.769	-0.94	-1.943	-2.255	11.616
11	49742778	49743850	TMSB10	7.648	7.167	9.889	9.446	11.55
4	21652784	21669638	LOC113891189	3.623	5.585	6.694	6.482	11.521
29	21999597	22007274	LOC113886272	NA	NA	-1.968	3.858	11.44
27	13989220	14142127	WWC2	6.889	6.405	7.257	5.288	11.32
10	79581547	79590358	NPC2	5.136	3.921	8.113	4.324	11.214
23	35777885	35790031	LOC113881912	NA	NA	NA	-3.133	11.21
29	11731065	11740940	LOC113886288	NA	NA	NA	NA	11.209
29	12860244	12875509	LOC113886290	NA	NA	NA	NA	11.14

Table 4.3: Significant tissue specific gene ontology pathways.

Liver	Term	Ont	N	Up	Down	P.Up	P
GO:0043167	ion binding	MF	16	3	13	0.992	0.035
GO:0072359	circulatory system development	BP	5	0	5	1	0.059
GO:0071704	organic substance metabolic process	BP	20	5	15	0.977	0.07
GO:0043169	cation binding	MF	14	3	11	0.979	0.078
GO:0046872	metal ion binding	MF	14	3	11	0.979	0.078
GO:0044238	primary metabolic process	BP	19	5	14	0.966	0.097

Muscle	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0071944	cell periphery	CC	12	11	1	0.04	0.996
GO:0016020	membrane	CC	19	16	3	0.053	0.987
GO:0016021	integral component of membrane	CC	11	10	1	0.059	0.993
GO:0031224	intrinsic component of membrane	CC	11	10	1	0.059	0.993
GO:0005886	plasma membrane	CC	11	10	1	0.059	0.993
GO:0005623	cell	CC	22	18	4	0.063	0.982
GO:0044464	cell part	CC	22	18	4	0.063	0.982
GO:0005576	extracellular region	CC	8	3	5	0.982	0.09
GO:0048584	positive regulation of response to stimulus	BP	4	1	3	0.981	0.095

Brain	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0016043	cellular component organization	BP	5	0	5	1	0.015
GO:0071840	cellular component organization or biogenesis	BP	5	0	5	1	0.015
GO:1901363	heterocyclic compound binding	MF	5	0	5	1	0.015
GO:0097159	organic cyclic compound binding	MF	5	0	5	1	0.015
GO:0022607	cellular component assembly	BP	4	0	4	1	0.036
GO:0044085	cellular component biogenesis	BP	4	0	4	1	0.036
GO:0005576	extracellular region	CC	4	4	0	0.092	1
GO:0044421	extracellular region part	CC	4	4	0	0.092	1
GO:0005488	binding	MF	16	6	10	0.97	0.093

Lung	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0070887	cellular response to chemical stimulus	BP	5	0	5	1	0.016
GO:0042221	response to chemical	BP	5	0	5	1	0.016
GO:0071495	cellular response to endogenous stimulus	BP	4	0	4	1	0.037
GO:0071310	cellular response to organic substance	BP	4	0	4	1	0.037
GO:0009719	response to endogenous stimulus	BP	4	0	4	1	0.037
GO:0010033	response to organic substance	BP	4	0	4	1	0.037
GO:0050789	regulation of biological process	BP	13	4	9	0.986	0.055
GO:0065007	biological regulation	BP	15	5	10	0.983	0.06
GO:0065009	regulation of molecular function	BP	6	1	5	0.994	0.061
GO:0009987	cellular process	BP	17	6	11	0.981	0.063

Continued on next page

Table 4.3 – continued from previous page

Placenta	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0005737	cytoplasm	CC	25	17	8	0.009	0.998
GO:0043168	anion binding	MF	5	5	0	0.016	1
GO:0036094	small molecule binding	MF	5	5	0	0.016	1
GO:0005622	intracellular	CC	28	18	10	0.017	0.995
GO:0044424	intracellular part	CC	28	18	10	0.017	0.995
GO:0044444	cytoplasmic part	CC	14	10	4	0.031	0.993
GO:0005515	protein binding	MF	14	10	4	0.031	0.993
GO:0097367	carbohydrate derivative binding	MF	4	4	0	0.036	1
GO:0097458	neuron part	CC	4	4	0	0.036	1
GO:1901265	nucleoside phosphate binding	MF	4	4	0	0.036	1
GO:0000166	nucleotide binding	MF	4	4	0	0.036	1
GO:0017076	purine nucleotide binding	MF	4	4	0	0.036	1
GO:0035639	purine ribonucleoside triphosphate binding	MF	4	4	0	0.036	1
GO:0032555	purine ribonucleotide binding	MF	4	4	0	0.036	1
GO:0032553	ribonucleotide binding	MF	4	4	0	0.036	1
GO:0065008	regulation of biological quality	BP	9	7	2	0.04	0.994
GO:0005623	cell	CC	34	20	14	0.042	0.983
GO:0044464	cell part	CC	34	20	14	0.042	0.983
GO:0003008	system process	BP	6	5	1	0.059	0.994
GO:0003824	catalytic activity	MF	17	11	6	0.061	0.981

5 General Discussion

Events that happen during fetal development are regulated by many genetic and epigenetic factors and affect adult phenotypes and lifelong health. During fetal development, the process of forming organs and tissues is mediated by tissue-specific patterns of gene expression. In this thesis I have addressed the patterns of expression at a key fetal stage to provide information on qualitative and quantitative changes in the transcriptome during normal development. This information contributes to our understanding of the mechanisms involved in the development of specific phenotypes. Many of the phenotypes are sex specific or differ between the breeds. Prior to this thesis, a substantial amount of research had been focused on sex and parental genome effects on both prenatal and postnatal cattle development (Xiang et al., 2013; Seo et al., 2016; Fang et al., 2019). The interpretation of these data is dependent on a well-assembled and annotated bovine sex chromosomes, which was lacking. The cattle genome assembly (ARS-UCD 1.2) which was used as the reference when I started this work was derived from a female and was thus missing the Y chromosome. With colleagues, we sequenced the genomes of the two cattle sub-species, *Bos taurus indicus* and *Bos taurus taurus*, from the lung of a crossbred fetus and assembled and annotated the haplotype resolved parental genomes (see § 6), and I then focused on the assembly and annotation of the sex chromosomes (see § 2).

The bovine sex chromosome sequences reported in this thesis have complete pseudoautosomal regions (PAR) and the Y chromosome has the three X-degenerate (X-d) regions fully assembled. The PAR comprises 31 genes and includes genes missing from the current cattle, sheep and goat reference genome sequences. We found a total of 16 paired X-Y gametologues, which are genes outside the PAR but present on both the X and Y, five of which are conserved between human, pig, horse and cattle. Although these cattle sex chromosome assemblies are the most complete to date, technical limitations mean that there are still some gaps, especially for the Y chromosome. The Y chromosome assembly is not full length but does contain complete PAR, X-d and ampliconic regions. However, the location of X-d2 and ampliconic regions in relation to the PAR remains to be established. The total length of the assembled Y chromosome sequence appears to be one

third of actual Y chromosome. Much of the missing sequence is probably the extensive heterochromatic regions that are currently very difficult to span, even with the latest long-read sequencing technologies. We placed these multicopy genes in the assembly based on their positions in RH and linkage maps, but the maps only have low resolution. In the future it may be possible to span these regions with a single sequence using ultra-long reads, now available using Pacific BioSciences or Oxford Nanopore technology.

Although the assembly and annotation of these cattle sex chromosomes are not complete, the precise identification of the PAR boundary and annotation of the gametologues enabled me to explore the differential gene expression between males and females and the correlation with sex-specific phenotypes (see section § 3). X chromosome inactivation (XCI) and dosage compensation has been observed in many mammals (Ohno et al., 1959; Lyon, 1961; Heard and Disteché, 2006) including cattle (De La Fuente et al., 1999; Xue et al., 2002). There have been several RNA-seq studies reporting X chromosome dosage compensation in early bovine embryos, germ cells and adult tissues (Ka et al., 2016; Duan et al., 2019). However, these studies have used reference genomes containing poorly assembled sex chromosomes assemblies (UMD3.1.1 and Btau 4.0) that are missing many genes, complicating interpretation of results.

My study addressed the X chromosome inactivation (XCI) pattern at mid-gestation in cattle. We found 24 non-PAR, X chromosome genes with significantly higher expression in females than in males. These genes may escape XCI. I found that all PAR genes also escape XCI with both alleles being expressed. The ratio of non-PAR X chromosome gene expression versus the autosomal gene expression in female showed no difference compared with the ratio in males, supporting Ohno's hypothesis (Lyon, 1961) that there is a balance in the overall X chromosome gene expression between sexes. Nevertheless, I found that the ratio of expression between the PAR and autosomal genes (PAR:A ratio) was significantly lower in females than in males, suggesting that XCI may extend into the PAR region. Interestingly, I found the expression of X-gametologues in the female was generally higher (approximately 1.2-1.8 times) than the expression in the male. This ratio differed among tissues, indicating that XCI escape was not the same in all tissues

and may be related to sex differences in tissue development (Deng et al., 2014). Both these observations are consistent with data from human adult tissue (Balaton et al., 2015; Tukiainen et al., 2017).

Notably, we found 24 candidate XCI escapees in the fetus, the majority of which were PAR genes and X-gametologues (located on the short arm). This is far fewer than the number of genes that are known to escape XCI in human adult tissues (Tukiainen et al., 2017). Therefore, I re-analysed published sex-specific expression data for cattle adult tissues (Seo et al., 2016) using our latest sex chromosome assemblies. I found that the number of X chromosome DE genes between female and male in adult tissues was ~6 times more than observed in fetal tissues, which is similar to the number of XCI escapes in human adult tissues. XCI is dynamic and the number of genes that escape XCI changes at different developmental stages, as shown in mouse and human (Marks et al., 2015; Xie et al., 2016). There are differences in XCI among species: in humans ~15-30% X-linked genes escape XCI (Balaton et al., 2015), but in mice only 3-7% of X-chromosome genes consistently escape XCI (Carrel and Brown, 2017).

In the current study, we use RNA-seq data to identify genes that are candidate XCI escapees, based on expression difference between sexes. However, we could not confirm which candidates are real XCI escapee as I could not determine the allelic contribution using short read data. It would be interesting to explore XCI escapees more fully in the bovine model at different developmental stages, using Iso-seq (long reads) which would enable the transcripts to be assigned to their chromosome of origin. In addition, single-cell data would enable us to investigate whether XCI is consistent among different cells in a tissue. A systematic analysis of the dynamics of XCI status at different developmental stages in each tissue would facilitate the more detailed exploration of the role of XCI in phenotypic variation between sexes.

In humans, some of the XCI escapee genes are clustered on the human X chromosome short arm and their distribution often coincides with topologically associating domains or TADs (Marks et al., 2015). It has been suggested that this clustering within chromatin regions is a reason why humans have more XCI genes than mice (Posynick and Brown,

2019). As my data suggest that cattle have a similar percentage of X- chromosome genes that escape XCI as humans, it would be interesting to compare the clustering patterns of XCI genes on X chromosome between human and cattle and whether these coincide with TADs. Although the Hi-C data we used to construct the genome assemblies could be used to explore TADs, the depth of coverage was not sufficient to define TADs on the sex chromosomes.

Sex differences in the transcriptome among all five tissues were mainly driven by six paired gametologues (*ZFX/ZFY*, *TXLNG/TXLNGY*, *SHROOM2/SHROOM2Y*, *KDM6A/UTY*, *ZRSR2/ZRSR2Y* and *EIF2S3/EIF2S3Y*), it has been argued that the expression level of critical regulatory genes involves regulation of both X and Y gametologues (Brockdorff and Turner, 2015). In my study I showed that the gametologues had significant differences in expression levels between sexes. When we combined the expression of X and Y gametologues in male and compared this with the combined expression from the 2 female X chromosomes, we found that some gametologues showed a much higher expression level in males. Interestingly, three X-gametologues (*USP9X*, *OFDI*, *DDX3X*) had the same level of expression between males and females, although it was the Y-gametologues that were highly expressed, indicating the possible male-specific function for the Y encoded gene at this developmental stage.

The ancestral X-Y gametologues are involved in transcriptional/translational regulation and chromatin modification (Bellott et al., 2014; Balaton et al., 2015). In our study, expression of the *ZFX/ZFY* transcription factor in fetal brain explained up to 80% of effects on sex related phenotypic difference. Although the sequence of *ZFX* and *ZFY* are similar, with 90% identity at the protein level, *ZFY* may have a different function than *ZFX*, resulting in the phenotypic differences. Further studies on how these ancestral X-Y gametologues regulate gene activity will contribute to our understanding of sexual dimorphism and differences between the sexes in health and disease traits.

Examining genome-wide gene expression in the Brahman and Angus fetuses (§ 4), we identified several genes consistently had a high level of expression in liver, muscle and placenta across all genotypes, the majority of which belonged to metabolic pathways,

such as galactose, starch and sucrose metabolism. We found groups of DEGs between Brahman and Angus that were members of these pathways. The expression of these DEGs was consistently higher in Angus than in Brahman. These findings suggest that metabolic processes are of importance during embryonic development, as may be expected. The lower metabolic rates in indicine cattle may help them to cope better with heat and environmental stresses (Hansen, 2004).

Parental genome effects on phenotypic traits have been of interest for a long time. The experimental design of my PhD enabled me to address the effects of genetic variation on fetal development, and also to explore parent of origin effects. Previous studies using the same experimental material have reported that the maternal genome significantly affects fetal liver, placenta, and muscle weights (Xiang et al., 2013; Xiang et al., 2014). In my study, I identified thousands of genes that are differentially expressed (DE) between Angus and Brahman in each of the 5 tissues studied. Few of the genes that were DE between purebred groups were DE in the reciprocal crosses, except for in the liver where there were many DE genes. This suggested that generally maternal and paternal derived alleles contributed equally to expression, irrespective of the parent of origin. However, there were some DEGs that were either parental genome driven or breed (Taurine or Indicine) driven in each tissue. These genes could be considered as potentially imprinted genes with unequal allelic expression. Using short reads, it was not possible to accurately assign many of the transcripts to parental genomes without parentally phased whole genome sequences genotypes, so the allelic contribution could not be confirmed.

The majority of DEGs between purebred groups had a similar level of expression in reciprocal-crosses, suggesting that the phenotypic differences (Xiang et al., 2013; Xiang et al., 2014) may be driven by a small group of DEGs, which have different expression levels between purebred groups and also are differentially expressed between reciprocal-crosses. These DEGs showed maternal, paternal or breed driven expression patterns and thus are of interest for further studies to understand the complex mechanisms that affect prenatal development and differences in phenotypes between indicine and taurine cattle.

Identification of differentially expressed genes in divergent beef cattle breeds and their

crosses could help to identify those genes that affect economically important production traits and environmental adaptation. Correlation between expression and variation in quantitative traits would strengthen the evidence for gene effects. For example, in a recently published Genome-Wide Association study on residual feed intake (RFI) and its component traits (Zhang et al., 2020), the SNP most significantly associated with metabolic body weight (MWT) and average daily gain (ADG) was downstream of *PLAG1* (Pleomorphic adenoma gene 1). In our dataset, *PLAG1* was more highly expressed in Angus than Brahman placenta. *PLAG1* regulates many genes, including growth factors such as insulin-like growth factor 2 (*IGF2*), and is a regulator of growth and reproduction (Juma et al., 2016).

To compare expression between breeds, I would have liked to remove the confounding effects of breed and sex in the analyses. However, because of the limited sample size (3 vs 3 or 2 vs 3), we decided to ignore sex and focus on breed differences. An analysis that took into account breed and sex, may help identify sex-linked paternal/maternal driven genes. This would further contribute to understanding the control of sex-specific phenotype differences.

We generated a full set of miRNA data (more than 50 million reads per sample) from liver and muscle, but time constraints meant these data have not been fully analysed. We have identified several highly expressed known miRNA such as bta-miR-143 and bta-let-7 in liver. In a study of feed efficiency in cattle, bta-miR-143 was found to have a higher level of expression in liver in high residual feed intake (RFI) cattle than low RFI cattle (Al-Husseini et al., 2016). In human and mouse, the let-7 miRNA family regulate the expression of growth factor gene *IGF2*. *IGF2* was highly expressed across five tissues in our RNA-seq dataset. Further studies of differentially expressed miRNA and comparing the predicted targets of these with known DEGs from the RNA-seq data would allow us to explore the relationship between miRNA and gene expression. The high coverage of these miRNA data also allows us to characterise miRNA profiling in tissues and to identify novel miRNAs that may play roles in regulating gene expression in bovine fetal development.

One interesting question I would have liked to address is the relationship between epi-

genetics (DNA methylation and histone modification) and gene activity. Combining the RNA-seq data with methylation and histone data would help identify potential regulatory elements. This would then make it possible to explore chromatin accessibility around DEGs. However, a lack resources and difficulty in generating ATAC-seq from frozen samples, meant that the data were not available.

Iso-seq data allows the allele specificity of expression to be tested. Some Iso-seq data was generated, but only from one of the hybrid fetuses. In collaboration with bioinformaticians from Pacific Biosciences, we phased the allelic isoforms in the hybrid fetus and tested several known imprinted genes. Results suggest Iso-Seq could be used to identify imprinted genes based on haplotype phasing results. Given these interesting results, data is now being generated from additional samples, which will help to confirm the status of the putatively imprinted genes identified here.

In summary, I studied sex-specific and breed-specific effects between *Bos taurus taurus* and *Bos taurus indicus* using the most complete cattle reference genome assemblies, which I helped create. The cattle sex chromosomes that I assembled identified the PAR, X-degenerate regions and the locations of gametologues, providing an invaluable reference for future sex-specific studies. The gametologues that were differentially expressed between sexes across five tissues contributed to fetal weight differences between sexes. The DEGs identified between pure breeds and reciprocal-cross with maternal/paternal expression patterns provide a reference point for further parent of origin study.

In this thesis I have shown that genomic and transcriptomic data can provide a new understanding of developmental processes. As the sequencing technologies improve and costs decline further, more types of “omics” data will be produced that will help us to better understand the biological complicity of embryonic development and prenatal sex-specific effects that create phenotypic difference between the sexes in indicine and taurine cattle.

References

- Balaton, B.P., Cotton, A.M., and Brown, C.J., 2015. Derivation of consensus inactivation status for x-linked genes from genome-wide studies. *Biol sex differ* [Online], 6, p.35.
- Bellott, D.W., Hughes, J.F., Skaletsky, H., Brown, L.G., Pyntikova, T., Cho, T.J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R.S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., Watt, J., Holder, M., Lee, S., Nazareth, L., Alfoldi, J., Rozen, S., Muzny, D.M., Warren, W.C., Gibbs, R.A., Wilson, R.K., and Page, D.C., 2014. Mammalian y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* [Online], 508(7497), pp.494–9.
- Brockdorff, N. and Turner, B.M., 2015. Dosage compensation in mammals. *Cold spring harbor perspectives in biology*, 7(3), p.a019406.
- Carrel, L. and Brown, C.J., 2017. When the lyon(ized chromosome) roars: ongoing expression from an inactive x chromosome. *Philos trans r soc lond b biol sci* [Online], 372(1733).
- De La Fuente, R., Hahnel, A., Basrur, P.K., and King, W.A., 1999. X inactive-specific transcript (xist) expression and x chromosome inactivation in the preattachment bovine embryo. *Biol reprod* [Online], 60(3), pp.769–75.
- Deng, X., Berletch, J.B., Nguyen, D.K., and Disteche, C.M., 2014. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat rev genet* [Online], 15(6), pp.367–78.
- Duan, J.E., Shi, W., Jue, N.K., Jiang, Z., Kuo, L., O'Neill, R., Wolf, E., Dong, H., Zheng, X., Chen, J., and Tian, X.C., 2019. Dosage compensation of the x chromosomes in bovine germline, early embryos, and somatic tissues. *Genome biol evol* [Online], 11(1), pp.242–252.
- Fang, L., Jiang, J., Li, B., Zhou, Y., Freebern, E., Vanraden, P.M., Cole, J.B., Liu, G.E., and Ma, L., 2019. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. *Communications biology* [Online], 2(1), p.100.

- Hansen, P., 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Animal reproduction science*, 82, pp.349–360.
- Heard, E. and Disteché, C.M., 2006. Dosage compensation in mammals: fine-tuning the expression of the x chromosome. *Genes dev* [Online], 20(14), pp.1848–67.
- Al-Husseini, W., Chen, Y., Gondro, C., Herd, R.M., Gibson, J.P., and Arthur, P.F., 2016. Characterization and profiling of liver micrnas by rna-sequencing in cattle divergently selected for residual feed intake. *Asian-australas j anim sci* [Online], 29(10), pp.1371–82.
- Juma, A.R., Dandimopoulou, P.E., Grommen, S.V., Van de Ven, W.J., and De Groef, B., 2016. Emerging role of *plag1* as a regulator of growth and reproduction. *J endocrinol* [Online], 228(2), R45–56.
- Ka, S., Ahn, H., Seo, M., Kim, H., Kim, J.N., and Lee, H.J., 2016. Status of dosage compensation of x chromosome in bovine genome. *Genetica* [Online], 144(4), pp.435–44.
- Lyon, M.F., 1961. Gene action in the x-chromosome of the mouse (*mus musculus* l.) *Nature* [Online], 190(4773), pp.372–373.
- Marks, H., Kerstens, H.H.D., Barakat, T.S., Splinter, E., Dirks, R.A.M., Mierlo, G. van, Joshi, O., Wang, S.-Y., Babak, T., Albers, C.A., Kalkan, T., Smith, A., Jouneau, A., Laat, W. de, Gribnau, J., and Stunnenberg, H.G., 2015. Dynamics of gene silencing during x inactivation using allele-specific rna-seq. *Genome biology* [Online], 16(1), p.149.
- Ohno, S., Kaplan, W.D., and Kinoshita, R., 1959. Formation of the sex chromatin by a single x-chromosome in liver cells of *rattus norvegicus*. *Exp cell res* [Online], 18, pp.415–8.
- Posynick, B.J. and Brown, C.J., 2019. Escape from x-chromosome inactivation: an evolutionary perspective. *Front cell dev biol* [Online], 7, p.241.

- Seo, M., Caetano-Anolles, K., Rodriguez-Zas, S., Ka, S., Jeong, J.Y., Park, S., Kim, M.J., Nho, W.G., Cho, S., Kim, H., and Lee, H.J., 2016. Comprehensive identification of sexually dimorphic genes in diverse cattle tissues using rna-seq. *Bmc genomics* [Online], 17, p.81.
- Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B.B., Castel, S.E., Karczewski, K.J., Aguet, F., Byrnes, A., Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E.B.I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U.o.C.S.C., Lapalainen, T., Regev, A., Ardlie, K.G., Hacohen, N., and MacArthur, D.G., 2017. Landscape of x chromosome inactivation across human tissues. *Nature* [Online], 550(7675), pp.244–248.
- Xiang, R., Ghanipoor-Samami, M., Johns, W.H., Eindorf, T., Rutley, D.L., Kruk, Z.A., Fitzsimmons, C.J., Thomsen, D.A., Roberts, C.T., and Burns, B.M., 2013. Maternal and paternal genomes differentially affect myofibre characteristics and muscle weights of bovine fetuses at midgestation. *Plos one*, 8(1).
- Xiang, R., Lee, A.M., Eindorf, T., Javadmanesh, A., Ghanipoor-Samami, M., Gugger, M., Fitzsimmons, C.J., Kruk, Z.A., Pitchford, W.S., and Leviton, A.J., 2014. Widespread differential maternal and paternal genome effects on fetal bone phenotype at mid-gestation. *Journal of bone and mineral research*, 29(11), pp.2392–2404.
- Xie, P., Ouyang, Q., Leng, L., Hu, L., Cheng, D., Tan, Y., Lu, G., and Lin, G., 2016. The dynamic changes of x chromosome inactivation during early culture of human embryonic stem cells. *Stem cell res* [Online], 17(1), pp.84–92.

Xue, F., Tian, X.C., Du, F., Kubota, C., Taneja, M., Dinnyes, A., Dai, Y., Levine, H., Pereira, L.V., and Yang, X., 2002. Aberrant patterns of x chromosome inactivation in bovine clones. *Nat genet* [Online], 31(2), pp.216–20.

Zhang, F., Wang, Y., Mukiibi, R., Chen, L., Vinsky, M., Plastow, G., Basarab, J., Stothard, P., and Li, C., 2020. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: i: feed efficiency and component traits. *Bmc genomics* [Online], 21(1), p.36.

6 Supporting Publication: Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle

Wai Yee Low¹, Rick Tearle¹, Ruijie Liu¹, Sergey Koren², Arang Rhie², Derek M. Bickhart³, Benjamin D. Rosen⁴, Zev N. Kronenberg⁵, Sarah B. Kingan⁶, Elizabeth Tseng⁶, Françoise Thibaud-Nissen⁷, Fergal J. Martin⁸, Konstantinos Billis⁸, Jay Ghurye⁹, Alex R. Hastie¹⁰, Joyce Lee¹⁰, Andy W. C. Pang¹⁰, Michael P. Heaton¹¹, Adam M. Phillippy², Stefan Hiendleder¹, Timothy P. L. Smith¹¹ & John L. Williams¹

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia, Australia

²Genomic Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

³Dairy Forage Research Center, ARS USDA, Madison, WI, USA

⁴Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, Maryland, USA

⁵Phase Genomics, 4000 Mason Road, Suite 225, Seattle, WA 98195, USA

⁶Pacific Biosciences, Menlo Park, CA 94025, USA

⁷National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

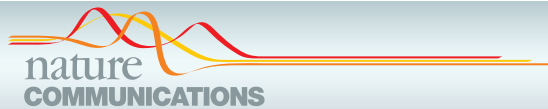
⁸European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

⁹Center for Bioinformatics and Computational Biology, Lab 3104A, Biomolecular Science Building, University of Maryland, College Park, Maryland, USA

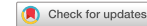
¹⁰ Genome Campus, Hinxton, Cambridge CB10 1SD, UK

¹¹US Meat Animal Research Centre, ARS USDA, Clay Centre, Nebraska, USA

Published in 2020, Nature Communications 11, 2071



ARTICLE


<https://doi.org/10.1038/s41467-020-15848-y>

OPEN

Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle

Wai Yee Low ¹, Rick Tearle¹, Ruijie Liu¹, Sergey Koren ², Arang Rhie ², Derek M. Bickhart³, Benjamin D. Rosen ⁴, Zev N. Kronenberg⁵, Sarah B. Kingan ⁶, Elizabeth Tseng⁶, Françoise Thibaud-Nissen ⁷, Fergal J. Martin ⁸, Konstantinos Billis ⁸, Jay Ghurye⁹, Alex R. Hastie¹⁰, Joyce Lee ¹⁰, Andy W. C. Pang ¹⁰, Michael P. Heaton ¹¹, Adam M. Phillippy ², Stefan Hiendleder ^{1✉}, Timothy P. L. Smith ^{11✉} & John L. Williams ^{1✉}

Inbred animals were historically chosen for genome analysis to circumvent assembly issues caused by haplotype variation but this resulted in a composite of the two genomes. Here we report a haplotype-aware scaffolding and polishing pipeline which was used to create haplotype-resolved, chromosome-level genome assemblies of Angus (taurine) and Brahman (indicine) cattle subspecies from contigs generated by the trio binning method. These assemblies reveal structural and copy number variants that differentiate the subspecies and that variant detection is sensitive to the specific reference genome chosen. Six genes with immune related functions have additional copies in the indicine compared with taurine lineage and an indicine-specific extra copy of fatty acid desaturase is under positive selection. The haplotyped genomes also enable transcripts to be phased to detect allele-specific expression. This work exemplifies the value of haplotype-resolved genomes to better explore evolutionary and functional variations.

¹The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia. ²Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA. ³Dairy Forage Research Center, ARS USDA, Madison, WI, USA. ⁴Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, MD, USA. ⁵Phase Genomics, 4000 Mason Road, Suite 225, Seattle, WA 98195, USA. ⁶Pacific Biosciences, Menlo Park, CA 94025, USA. ⁷National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. ⁸European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ⁹Center for Bioinformatics and Computational Biology, Lab 3104A, Biomolecular Science Building, University of Maryland, College Park, MD 20742, USA. ¹⁰Bionano Genomics, San Diego, CA, USA. ¹¹US Meat Animal Research Center, ARS USDA, Clay Center, NE, USA. ✉email: stefan.hiendleder@adelaide.edu.au; tim.smith2@usda.gov; john.williams01@adelaide.edu.au

About 10,000 years ago, cattle were domesticated from the aurochs, which ranged across Eurasia and North Africa but are now extinct¹. Modern day cattle belong to two subspecies, the humped zebu or indicine breeds (*Bos taurus indicus*) and the humpless taurine breeds (*Bos taurus taurus*), which arose from independent domestication events of genetically distinct aurochs populations².

During the past century, taurine breeds have been intensively selected for production traits, particularly milk and meat yield, and generally have higher fertility than indicine breeds. European taurine breeds, such as Angus, have excellent carcass and meat quality, high fertility, and reach puberty early. These breeds have been imported by farmers around the world to improve or replace less-productive breeds. However, while European taurine animals are well adapted to temperate environments, they do not thrive in hot, humid tropical environments with high disease and parasite challenge.

Indicine breeds originated from the Indus valley and later spread to Africa and across southeast Asia³. Between 1854 and 1926, the four indicine breeds, Ongole, Krishna, Gir, and Gujarat, were imported into the United States and crossed with European taurine cattle to create the Brahman breed. Current US Brahman cattle retain ~10% of their genome of taurine origin⁴. Brahman have a short, thick, glossy coat that reflects sunlight and loose skin that increases the body surface area exposed for cooling. While Brahman are less productive and have lower fertility than taurine breeds, they have desirable traits, such as heat tolerance, lower susceptibility to parasites such as ticks, and are more disease and drought resistant⁵.

We previously demonstrated a trio binning approach to assemble haplotypes of diploid individuals at the contig level. The quality of the contigs exceeded those of the best livestock reference genomes⁶. Here we present chromosome-level taurine (Angus) and indicine (Brahman) cattle genomes from a single crossbred individual that were assembled with haplotype-aware methodology that is less laborious than sequencing haploid clones⁷. The contiguity and accuracy of the final haplotype-resolved cattle assemblies set a high standard for diploid genomes and enable precise identification of genetic variants, from single-nucleotide polymorphisms (SNPs) to large structural variants (SVs). A further benefit of haplotype-resolved genomes is that they can be used to better interpret allele-specific expression in diploid transcriptome profiles. We identify allele-specific and novel transcripts using PacBio Iso-Seq reads mapped onto the haplotype-resolved genomes. Considering the large differences in production and adaptation traits between taurine and indicine cattle, comparison of genomes between the breeds will contribute to unveiling the mechanisms behind phenotypic differences among cattle including environmental adaptation, which is of substantial scientific and economic interest.

Results

De novo assembly and annotation of Angus and Brahman genomes. The initial creation of haplotigs (haplotype-specific contigs) was presented in the description of the trio binning method implemented in TrioCanu⁶. Briefly, a male *Bos taurus* hybrid fetus, from an Angus sire and a Brahman dam, was sequenced to ~136× long-read coverage, and the reads were sorted into parental haplotype bins based on *k*-mers that are unique to either the paternal or maternal genome, which were identified by short-read sequencing of the parents prior to assembly with TrioCanu. The initial assemblies comprised 1747 Angus haplotigs and 1585 Brahman haplotigs (Table 1). The haplotig N50 was 29.4 and 23.4 Mb for the Angus and Brahman, respectively.

For the present study, additional data were generated for the same hybrid fetus, including ~12× Hi-C reads, ~167× Bionano

optical map, and ~84× Illumina paired-end reads (Fig. 1), to provide haplotype-resolved scaffolding and identify assembly errors. Following haplotig assembly, two sets of scaffolds, one based on Hi-C and the other on optical map data, were generated for each haplotype. Three different scaffolding programs (3D-DNA, Proximo, and SALSA2) were evaluated using the Hi-C data (Supplementary Note 1). SALSA2 was found to be the best scaffolder and produced the closest agreement with the latest cattle reference ARS-UCD1.2. The scaffold N50 produced by SALSA2 was larger than that generated by optical map scaffolding, but the latter detected chimeric haplotig more accurately (i.e., a haplotig incorrectly assembled), which resulted in 29 and 36 breaks in the Angus and Brahman haplotigs, respectively (Supplementary Note 2). These chimeric breaks corrected four inter-chromosomal fusions in the initial haplotigs, involving two Brahman chromosomes (13 and 15) and six Angus chromosomes (8, 9, 12, 20, 23, 28).

After validation against a recombination map, gap filling, and error correction, the final assemblies, UOA_Angus_1 and UOA_Brahman_1, had chromosome sizes similar to the current cattle reference, ARS-UCD1.2 (Supplementary Fig. 1). Unlike some of the recent PacBio-based assemblies^{8,9}, which required an additional polishing step with Illumina short reads to correct the high indel error rates, the haplotype-resolved assemblies only required correction of a very small number of coding sequences, showing that polishing with short reads was unnecessary (Supplementary Note 3).

The Brahman genome was annotated by Ensembl and NCBI, whereas the Angus genome was annotated only by Ensembl (Supplementary Notes 3 and 4). A comparison of annotation features between the Angus, Brahman, and Hereford reference genomes is given in Supplementary Table 1. As the Ensembl pipeline was used to annotate all three cattle genomes, interpretation of results reported here used Ensembl release 96.

Assembly benchmarking and sequence contiguity assessments.

The per-base substitution quality values (QVs) for the UOA_Angus_1 and UOA_Brahman_1 reference assemblies were 44.63 and 46.38, respectively (Supplementary Table 2, Supplementary Note 5). The QV represents the phred-scaled probability of an incorrect base substitution in the assembly, hence these QVs indicate that the assemblies are >99.99% accurate at single base level. This is similar to the latest water buffalo assembly UOA_WB_1 (QV 41.96) and surpasses the recent goat ARS1 assembly (QV 34.5) by an order of magnitude. The Angus and Brahman assemblies had ~93% BUSCO completeness score, which demonstrates a high-quality (HQ) assembly of genes (Supplementary Table 3).

The Angus and Brahman assemblies have few gaps compared to most existing mammalian reference assemblies and are comparable to the human GRCh38, the latest Hereford cattle ARS-UCD1.2, and the water buffalo UOA_WB_1 reference genomes (Fig. 2a). For example, the Angus chromosome 24 was assembled without gaps. In terms of contiguity, these cattle reference genomes are comparable to the recent water buffalo UOA_WB_1 assembly⁹, which is the most contiguous ruminant genome published to date with <1000 contigs (Supplementary Fig. 2), although it is not fully haplotype resolved. While the cattle autosomes showed excellent contiguity, the Brahman X and Angus Y chromosomes were interrupted by 91 and 69 gaps, respectively.

Resolution of longer repeats. The use of long PacBio reads substantially improved repeat resolution compared with the previous cattle assembly UMD3.1.1, which was assembled from Sanger sequences¹⁰ (Fig. 2b). Approximately 49% of both Angus and

Table 1 Assembly statistics.

Breed	Assembly	Software	Assembly level	Number of sequences ^a	Number of gaps	N50 (Mb)	Assembly size (Gb)
Angus	PacBio	CANU	Haplotig	1747	0	29.4	2.6
Angus	PacBio+Hi-C	SALSA2	Scaffold	1515	235	104.6	2.6
Angus	PacBio+Optical map	Bionano Access	Scaffold	1595	181	35.2	2.6
Angus	UOA_Angus_1	PBJelly, Aarow, custom scripts	Chromosome	1435	277	102.8	2.6
Brahman	PacBio	CANU	Haplotig	1585	0	23.4	2.7
Brahman	PacBio+Hi-C	SALSA2	Scaffold	1370	216	72.6	2.7
Brahman	PacBio+Optical map	Bionano Access	Scaffold	1353	268	31.7	2.7
Brahman	UOA_Brahman_1	PBJelly, Aarow, custom scripts	Chromosome	1251	302	104.5	2.7

^aThere are 1405 and 1220 unplaced haplotigs in the final chromosome-level Angus and Brahman assemblies, respectively. These unplaced haplotigs comprise ~3.8% of total bases in the Angus assembly and ~2.1% of total bases in the Brahman assembly. Only the Brahman assembly has a complete mitochondrion sequence.

Brahman assemblies consist of repeat elements, which is consistent with other published mammalian assemblies, including human GRCh38, Hereford cattle ARS-UCD1.2, water buffalo UOA_WB_1, and goat ARS1. The two largest repeat families identified were Long Interspersed Nuclear Element (LINE) L1 and LINE/RTE-BovB, which covered ~25% of the chromosomes in both cattle sub-species (Supplementary Fig. 3). Satellite or centromeric repeats (>10 kb) accounted for 21% and 14% of repeats in unplaced scaffolds of Angus and Brahman, respectively. The 7% higher satellite and centromeric repeats in Angus unplaced scaffolds may be due to the presence of the Y chromosome in the Angus haplotype. The combination of the three most frequent repeat families, LINE L1, LINE/RTE-BovB, and satellite/centromeric repeats, covered ~40% of all unplaced bases, and repeat sequences were most frequently responsible for breaking sequence contiguity. The three cattle assemblies constructed using PacBio long reads that resolved repeats >2.5 kb, UOA_Angus_1, UOA_Brahman_1, and ARS-UCD1.2, provide significant improvements in repeat resolution over the previous Sanger-based cattle assembly (UMD3.1.1) (Fig. 2b). In both the Brahman and Angus assemblies, 20 out of 29 of the autosomes contained centromeric repeats within 100 kb of chromosome ends. Vertebrate telomeric repeats (TTAGGG)_n were found within 1 Mb of the ends of six Angus and five Brahman chromosomes. This demonstrates that some scaffolds approach chromosome-level assembly.

Discovery of indicus-specific fatty acid desaturase 2. One of the most diverged genomic regions between Brahman and Angus was observed on chromosome 15 (Fig. 3a). A region of ~1.4 Mb has three copies of fatty acid desaturase 2-like genes (*FADS2P1*) in Brahman, whereas the homologous region in the Angus only has two *FADS2P1* genes (Fig. 3b, c). In both Brahman and Angus, the *FADS2P1* genes are encoded by 10–12 exons, and the entire region was assembled completely without gaps for both genomes. The region also contains six genes annotated as olfactory receptor-like, with unknown functions, which had differences in their predicted gene models between Brahman and Hereford assemblies. Within the ~1.4 Mb region, there is a high level of sequence divergence for ~200 kb, which is where an extra copy of *FADS2P1* lies in Brahman. Searches for *FADS2P1* in other ruminant species with HQ genome assemblies revealed that only Brahman has three copies of the gene. The additional *FADS2P1* gene is ~53 kb long and is flanked by two other conserved *FADS2P1* genes. Searching whole-genome sequencing (WGS) short-read sequences from 38 animals used in this study showed that only Brahman animals had the extra copy, which was not present in any of the taurine individuals. (Supplementary Fig. 4). Considering that the Brahman genome is derived from four indicine breeds, the extra *FADS2P1* is likely a *Box taurus*

indicus-specific gene. We used a maximum likelihood-based estimate of ratio of non-synonymous (amino acid changes) to synonymous (silent changes) substitutions as implemented in CODEML to search for positively selected amino acid residues in *FADS2P1* and identified 16 significant positively selected sites, 10 of which are located in a small exon 7 of only 60 bp (Fig. 3d, Supplementary Table 4).

SNP and INDEL differences between Brahman and Angus.

Mapping short reads from Brahman and Angus to both reference genomes, UOA_Brahman_1 and UOA_Angus_1, revealed that the use of breed-specific reference genomes gave a lower count of all classes of genetic variants. Using WGS short reads from 5 Brahman and 6 Angus individuals, we identified ~24 million Brahman SNPs and ~11 million Angus SNPs, which were annotated using their own reference genome (Table 2, Supplementary Table 6). There were about twice as many INDELS in the Brahman (2,804,421 bp) than the Angus (1,381,548 bp) samples. Lower counts of SNPs, INDELS, and the four classes of SVs (i.e., BND, DEL, DUP, INV) were identified when the appropriate reference genome was used. For example, ~4% fewer SNPs were observed when Brahman individuals were mapped onto the Brahman instead of the Angus reference genome. Additional information on the use of SNPs for the analysis of selective sweeps in cattle is given in Supplementary Note 6.

SV differences between Brahman and Angus. We assessed the structural continuity of our Brahman and Angus genome assemblies against the current cattle reference genome assembly, ARS-UCD1.2, and against WGS datasets from 38 animals representing seven breeds, to ascertain the benefit of using haplotype-resolved assemblies for variant calling. To assess SV differences between Brahman and Angus and the cattle reference genomes, the haplotype-resolved assemblies were aligned to the ARS-UCD1.2 reference (Hereford). This detected insertions, deletions, tandem expansions, tandem contractions, repeat expansions, and repeat contractions¹¹ (Supplementary Fig. 5). Both tandem expansion/contraction and repeat expansion/contraction are repeat-type SVs. Detection of SVs was limited to sizes of 50–10,000 bp, and the total bp affected by SVs in Angus and Brahman were 10.9 and 21.8 Mb. This translates to approximately 0.4% and 0.8% of the Angus and Brahman genomes, respectively. Among the six classes of SVs examined, insertion/deletion types were the most prevalent in both Brahman and Angus genomes compared to ARS-UCD1.2.

We extracted Brahman- and Angus-specific SVs to study their distribution in genic and intergenic regions (Fig. 4a). Brahman-specific insertions/deletions overlapped ~4% of all genes, whereas

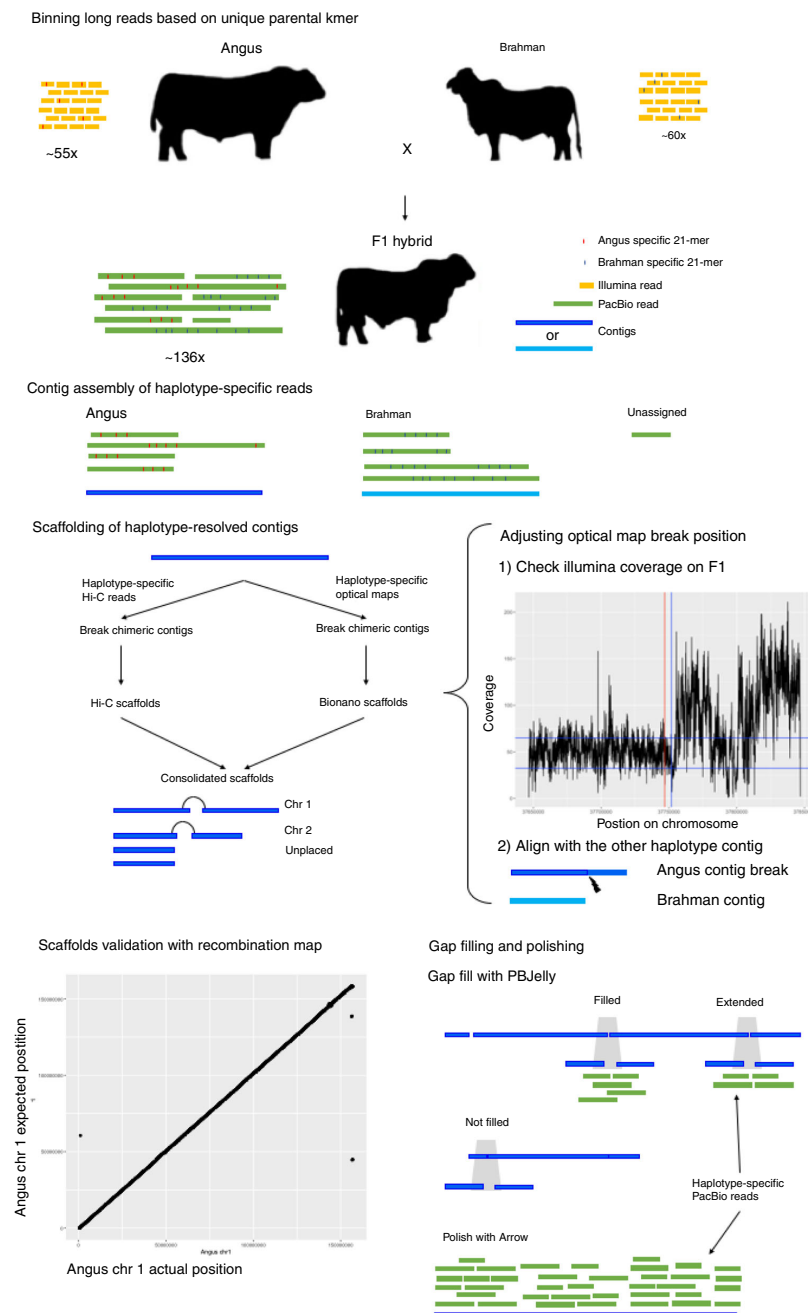


Fig. 1 An overview of the assembly methods. Long PacBio reads were binned to the respective haplotypes using parental-specific *k*-mers and unassigned reads were discarded. TrioCanu was used to assemble sequences from each haplotype into haplotigs. Each set of haplotigs was scaffolded separately with both Hi-C and optical map data (illustrated only for the Angus). Optical map breakpoints were accepted but are imprecise. Therefore, breakpoint positions were improved by observing if there are local drops in short-read coverage and/or where there is a break sequence alignment with the alternative haplotig. Hi-C and optical map-based scaffolds were checked for consistency and combined as a single set of scaffolds. Cattle recombination maps were used to validate the assembly. Each point on the scatter plot is the actual recombination marker coordinate on the latest reference genome and the expected position based on previous reference genome, UMD3.1. Finally, haplotype-specific marker reads were used to fill gaps and polish the sequence.

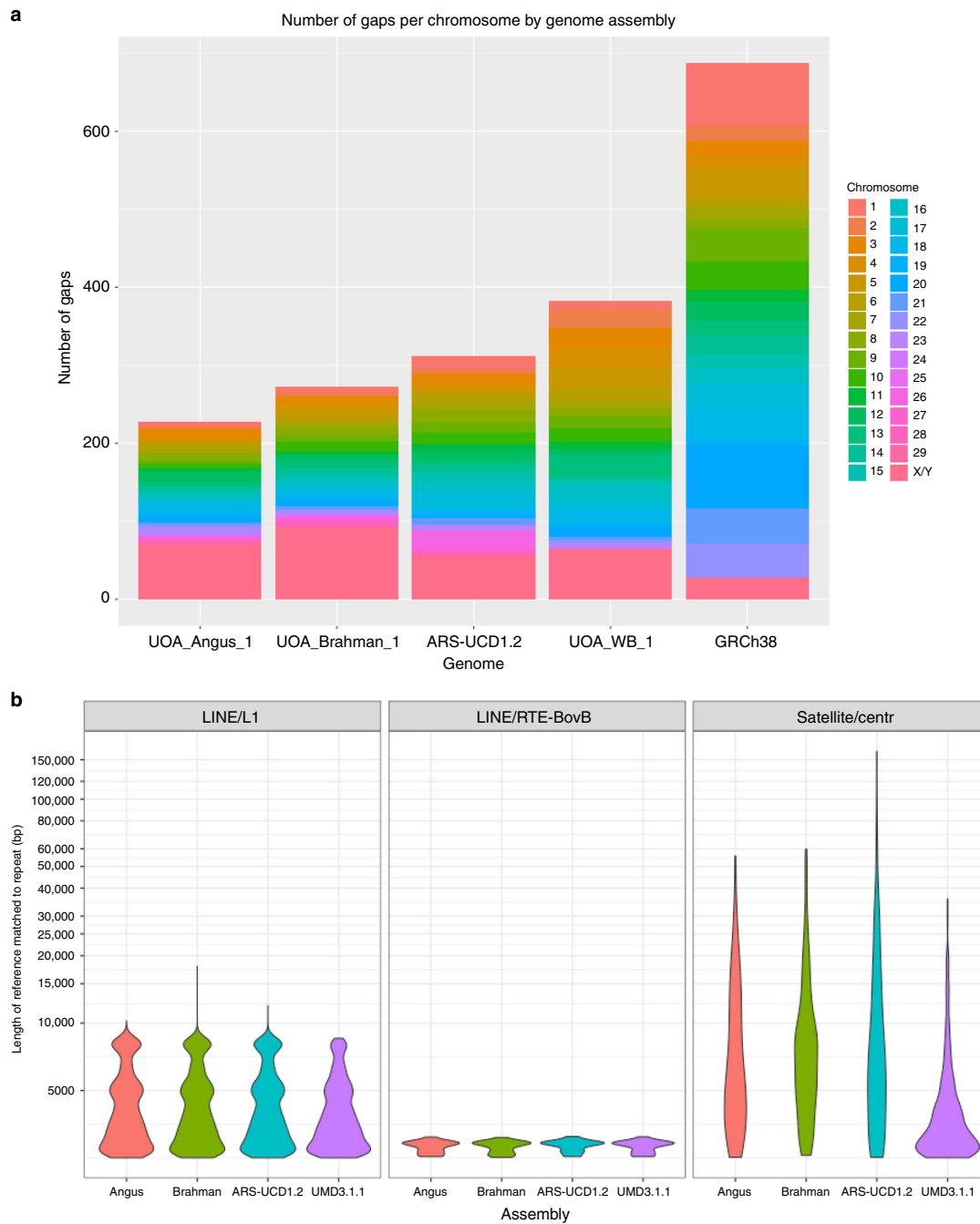


Fig. 2 Sequence contiguity and resolution of repeats. **a** Barplot of the number of gaps by chromosomes between various mammalian assemblies. **b** Violin plot of repeat families filtered for those >2.5 kb for LINE/L1, LINE/RTE-BovB, and satellite/centromeric repeats.

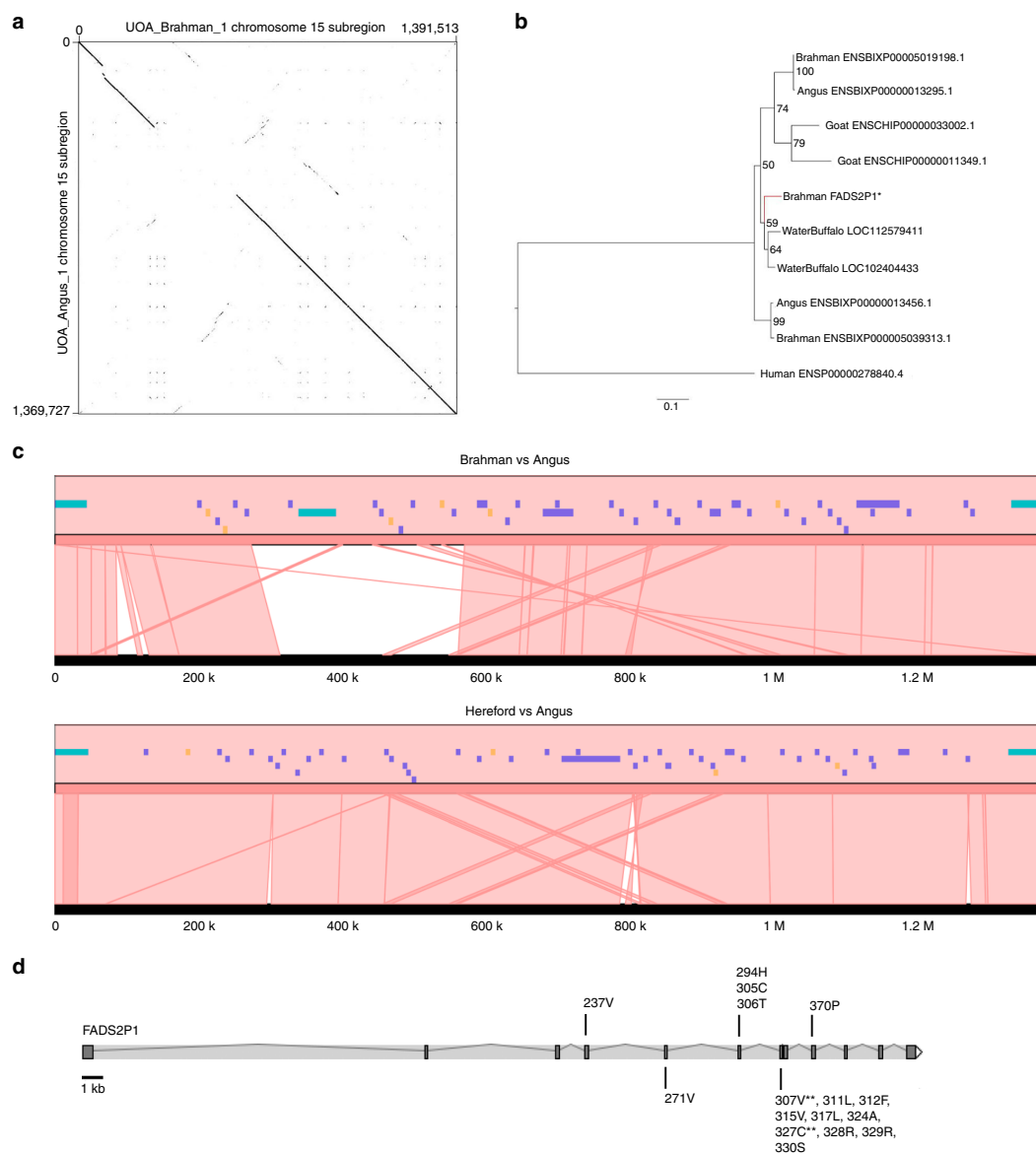


Fig. 3 Divergence of the *FADS2P1* locus between indicine and taurine cattle. **a** Dot plot of Brahman chromosome 15 between positions 3,748,952 to 5,140,465 against the homologous Angus chromosome between positions 78,799,177 to 80,168,904. The Brahman sequence was reverse complemented in the plot. **b** Maximum likelihood tree with 1000 bootstraps of *FADS2P1* homologous protein sequences. The extra Brahman *FADS2P1* copy is highlighted with asterisk (*) and its branch colored red. **c** Microsynteny plot showing a lack of sequence conservation between indicine and taurine breeds around the indicine-specific *FADS2P1* gene. All *FADS2P1* genes are colored turquoise, other genes purple, and pseudogenes orange. The upper plot compares Brahman to Angus and the lower plot compares Hereford to Angus. The track in black in both panels is the Angus reference. The Brahman *FADS2P1* gene Ensembl IDs are ENSBIXG00005007613, ENSBIXG00005021668, and ENSBIXG00005022680, whereas the Angus IDs are ENSBIXG00000018262 and ENSBIXG00000018381. The indicine-specific copy of *FADS2P1* is ENSBIXG00005007613. **d** Mapping of 16 positively selected sites onto the exons of Brahman *FADS2P1*. The residues with double asterisks (**) indicate they have $\text{Prob}(\omega > 1) > 0.99$ (i.e., highly significant positively selected sites).

Angus-specific insertions/deletions overlapped only ~1–2% of genes. Each repeat-type SV overlapped ~1% of genes in Brahman and <1% in Angus. The majority of SVs were found in intergenic regions, and when they overlapped with genes, they were

generally localized within introns. Over-representation of Gene Ontology (GO) terms was detected for Angus-specific insertions and tandem contractions and Brahman-specific insertion/deletion SVs at false discovery rate (FDR)-adjusted, Fisher's exact test,

Table 2 Polymorphism statistics.

Breed ^a	Reference	SNP	INDEL	BND	DEL	DUP	INV
Angus	Angus	10,615,122	1,381,548	38	84	22	3
Brahman	Angus	24,930,357	2,928,526	311	641	86	22
Angus	Brahman	16,504,067	2,090,735	159	182	40	11
Brahman	Brahman	23,876,357	2,804,421	279	481	97	18

BND complex structural variant, DEL deletion, DUP duplication, INV inversion.
^aBreed here refers to the ~10× WGS short reads used to align to the reference. BND, DEL, DUP, and INV are structural variant types called in Lumpy.

P value < 0.05 (Supplementary Table 7). No over-representation of GO terms was detected for any of the other breed-specific SV types. Interestingly, Brahman-specific insertion SVs have between 3- and 5.7-fold enrichment in phospholipid translocation (GO:0045332), lipid translocation (GO:0034204), lipid transport (GO:0006869), and lipid localization (GO:0010876) GO classes, which suggests that lipid distribution was most impacted by SVs.

Using WGS reads from different datasets, we identified subspecies-specific copy number variations (CNVs) that were masked by the absence or poorer resolution of sequence in the ARS-UCD1.2 reference. The input dataset for these analyses came from ~10× WGS short reads of 38 animals representing 7 cattle breeds. Each set of reads was aligned to all three reference genome assemblies (Hereford, Brahman, and Angus) and processed with SV callers designed to detect read depth differences and paired-end/split-read (PE) discordancy, respectively. The read-depth variation approach included the use of the V_{st} statistic^{12, 13} to identify genes with CNV between taurine or indicine lineages using the Brahman, Angus, or ARS-UCD1.2 assemblies (Fig. 4b, Supplementary Fig. 6). The values of V_{st} varied greatly depending on which reference genome was used for alignment, with taurine-based reference assemblies showing higher variance in copy number between the taurine and indicine lineage datasets (Supplementary Fig. 6) than the Brahman reference (Fig. 4b). Only autosomes were considered. Six CNV genes were found in Brahman, whereas four and eight CNV genes were found in Angus and Hereford, respectively (Fig. 5a–c). Prediction of CNV genes was sensitive to the assembly chosen, e.g., only *TMPRSS11D* and beta-defensin-like precursor were found to be copy number variable in more than one assembly. Among the 18 CNV genes differentiating indicine from taurine genomes, six unique gene families were identified, which were beta defensin, workshop cluster, trypsin-like serine protease, T cell receptor alpha chain, tachykinin receptor, and interferon-induced very large GTPase, all of which have immune-related functions. All of the CNV genes from these six families showed higher copy number in the indicine cattle lineage regardless of the assembly used. Intersection of liftover CNV regions (CNVRs) called using the Brahman assembly with repetitive elements on ARS-UCD1.2 showed a higher prevalence of CNVs that may have resulted from repeat expansion/contraction in the Brahman reference (1813) than in ARS-UCD1.2 (1238) or the Angus (1164) assemblies. FRC_align statistics showed a higher count of COMPR_PE and STECH_PE events in the Angus assembly (319 and 101, respectively) than the Brahman assembly (263 and 87, respectively), supporting the hypothesis that expansion and contraction of genomic sequence in the Angus assembly is the likely reason for these discrepancies. An olfactory receptor, two long non-coding RNAs and one putative protein, FAM90A12P, also had higher copy numbers among indicine animals. In contrast, ubiquitin-conjugating enzyme E2D3 and two keratin-associated protein 9 genes (*KRTAP9-1*, *KRTAP9-2*) had higher copy numbers in the taurine lineage.

We quantified the effects of using different reference assemblies for PE SV discovery. All SV calls of this type were converted into Hereford coordinates to facilitate comparisons. We removed 17, 9, and 18 PE SVs of all types from the Brahman, Angus, and Hereford assemblies that were likely false positives, as they were >1 Mb and did not correspond to aberrant read depth signal to support their SV calls. On average, 0.5% of each cattle genome was covered by CNVRs (Fig. 5d). The majority of CNVRs (at least 76% from each assembly) were found to be unique to one assembly. Among the Brahman CNVRs, only 10% intersected with Angus CNVRs, which suggests mis-assembly in the Hereford reference potentially due to compression of repetitive elements that are more difficult to resolve without phasing haplotypes using the trio binning method.

Allele-specific transcripts in haplotype-resolved genomes.

Among the PacBio error corrected Iso-Seq (circular consensus sequence (CCS)) reads pooled from seven tissues of the F1 hybrid fetus, 3,275,676 reads (55%) were classified as full-length non-concatamer (FLNC) reads. After processing with the isoseq3 software, 193,974 full-length, HQ consensus transcripts were generated. We mapped the HQ transcripts to the Brahman reference and obtained 99,329 uniquely mapped transcripts covering 20,940 non-overlapping loci representing 19,403 genes. Of these 99,329 Iso-Seq transcripts, 20,708 (20.8%) had a perfect exon-by-exon match to the reference annotation while 11,359 (11.4%) matched a reference transcript but was missing one or more of the 5' exons (indicator of 5' degradation or alternative start site). The majority of the remaining transcripts (59,158, 60%) were novel isoforms of known genes, with the remainder 6.8% of transcripts categorized as intergenic, genomic, or anti-sense that are likely cDNA artifacts. At the gene level, 13,754 of the 19,403 (71%) genes are annotated reference genes. Using the SQANTI2 transcript characterization tool, 83% of the Iso-Seq transcripts fell into coding regions of the Brahman annotation (Fig. 6a). The transcript length distribution ranged from 85 to 11,872 bp, with a median of 3853 bp and a mode of ~4 kb (Fig. 6b).

We validated the IsoPhase SNPs using (1) SNPs called from RNA-Seq data of the brain, liver, lung, muscle, and placenta of the F1 hybrid and (2) Angus SNPs derived from mapping Illumina WGS short reads of the F1 hybrid to the Brahman reference. As the RNA-Seq had greater coverage than Iso-Seq and the SNPs called from genomic DNA included non-transcribed regions, only SNPs that were in positions covered by at least 40 full-length Iso-Seq reads were retained. The concordance of filtered SNPs called from WGS, RNA-Seq, and Iso-Seq is very high (87%) (Fig. 6c). Of the 45,313 SNPs called by IsoPhase, 39,452 (87%) were validated by SNPs from RNA-Seq and WGS, whereas 876 (1.9%) were only validated by RNA-Seq and 2155 (4.7%) were only validated by WGS (Supplementary Note 7). SNP calls that showed inconsistencies could often be explained by lower Iso-Seq coverage, SNPs in homopolymer regions, or alignment artifacts.

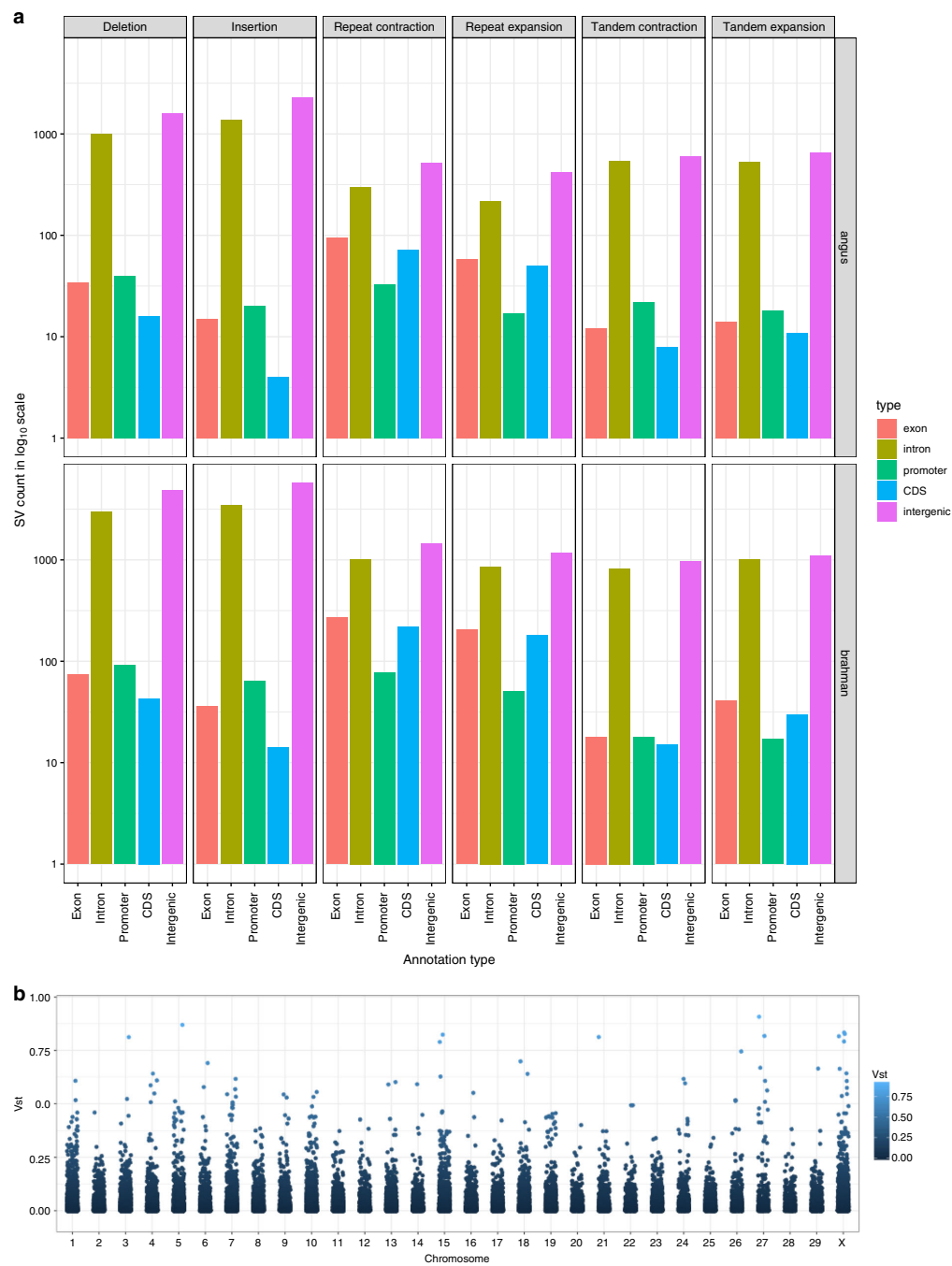


Fig. 4 Comparison of structural variants between Brahman and Angus. a Count in \log_{10} scale of 6 classes of SVs when overlapped with various annotation types. **b** Population differentiation for copy number variation as estimated by V_{st} along each chromosome for the taurine and indicine comparison using UOA_Brahman_1 as the reference.

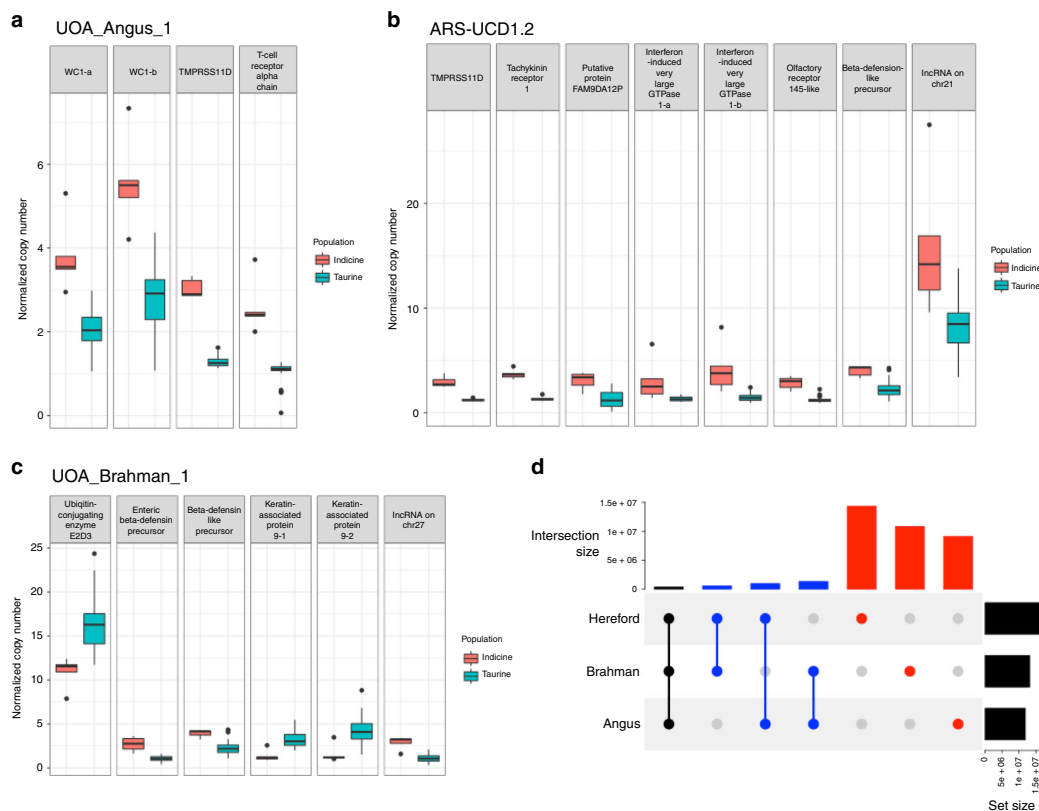


Fig. 5 Boxplot of normalized copy number of autosomal genes with $V_{st} > 0.3$. Only those CNV genes with average copy number difference of at least 1.5 copies between the taurine and indicine groups are shown. Dot plots of individual values are overlaid on top of boxplots to show minima and maxima as circles. The bounds of box show the 25th and 75th percentile, with the median drawn as a thick line between these two quartiles. The reference genomes were **a** UOA_Angus_1, **b** ARS-UCD1.2, and **c** UOA_Brahman_1. **d** Liftover of CNV regions from Brahman and Angus to Hereford ARS-UCD1.2 common coordinate for an assessment of intersection between them at base-pair resolution.

Our haplotype-resolved genomes allowed us to explore genes with allelic imbalance in expression. To assess allelic imbalance, the proportion of an allele from each breed was calculated as the normalized count of the Brahman allele divided by the sum of normalized counts of both Brahman and Angus alleles. All tissues showed evidence of imbalance in allelic expression (Shapiro test, P value < 0.01), which was most pronounced for liver, lung, muscle and placenta, whereas brain, heart and kidney were less affected (Fig. 6d). However, as the mammalian brain consists of a wide range of cell types and hence transcriptional complexity, brain tissue was chosen to demonstrate the phasing of transcripts to explore allele-specific expression. The most highly expressed Angus gene with allelic imbalance (ratio of 8 Angus:1 Brahman) in the brain was *ARIH2* (also known as *TRIAD1*), which is known to play a role in protein degradation via Cullin-RING E3 ubiquitin ligases¹⁴ (Fig. 6e, f). *ARIH2* expression in the liver, lung, muscle, and placenta was also higher from the Angus allele than the Brahman or maternal allele. The HQ transcripts included 23 different transcript isoforms of *ARIH2*; however, 66% of transcripts for this gene across the 7 tissues were represented by only 3 isoforms. The annotated exons of this gene were in good agreement with the RNA-Seq data (Supplementary Fig. 7).

The most highly expressed Brahman gene with allelic imbalance (ratio of 1 Angus:6 Brahman) in the brain was

Calmodulin (*CaM*), a heat-stable Ca^{2+} -binding protein that mediates the control of numerous physiological processes, including metabolic homeostasis, phospholipid turnover, ion transport, osmotic control, and apoptosis¹⁵ (Supplementary Fig. 8). Surprisingly, we also found allelic imbalance (ratio of 1 Angus:16.5 Brahman) in pregnancy-associated glycoprotein 1 (*PAG1*) with a higher expression of the Brahman allele in the brain and placenta but undetectable in other tissues. This gene was previously thought to be placenta specific and is used as a biomarker for embryo survival¹⁶.

Discussion

Traditional genome assembly approaches collapse haplotypes and therefore do not allow accurate assembly or the study of divergent, heterozygous regions. Here we demonstrate an assembly approach that yielded highly contiguous, haplotype-resolved Brahman and Angus cattle genomes from an F_1 hybrid of the two subspecies. Our analyses demonstrated that previous studies^{4, 17}, which mapped indicine sequences onto the taurine reference UMD3.1.1, will have identified loci where the subspecies are fixed for different alleles. Calling SNPs in transcripts from a diploid hybrid with both haplotypes decoded provides accurately phased transcripts for studies on the role of allele-specific expression in, e.g., hybrid vigor or heterosis. The phasing of Iso-Seq transcripts in reciprocal crosses

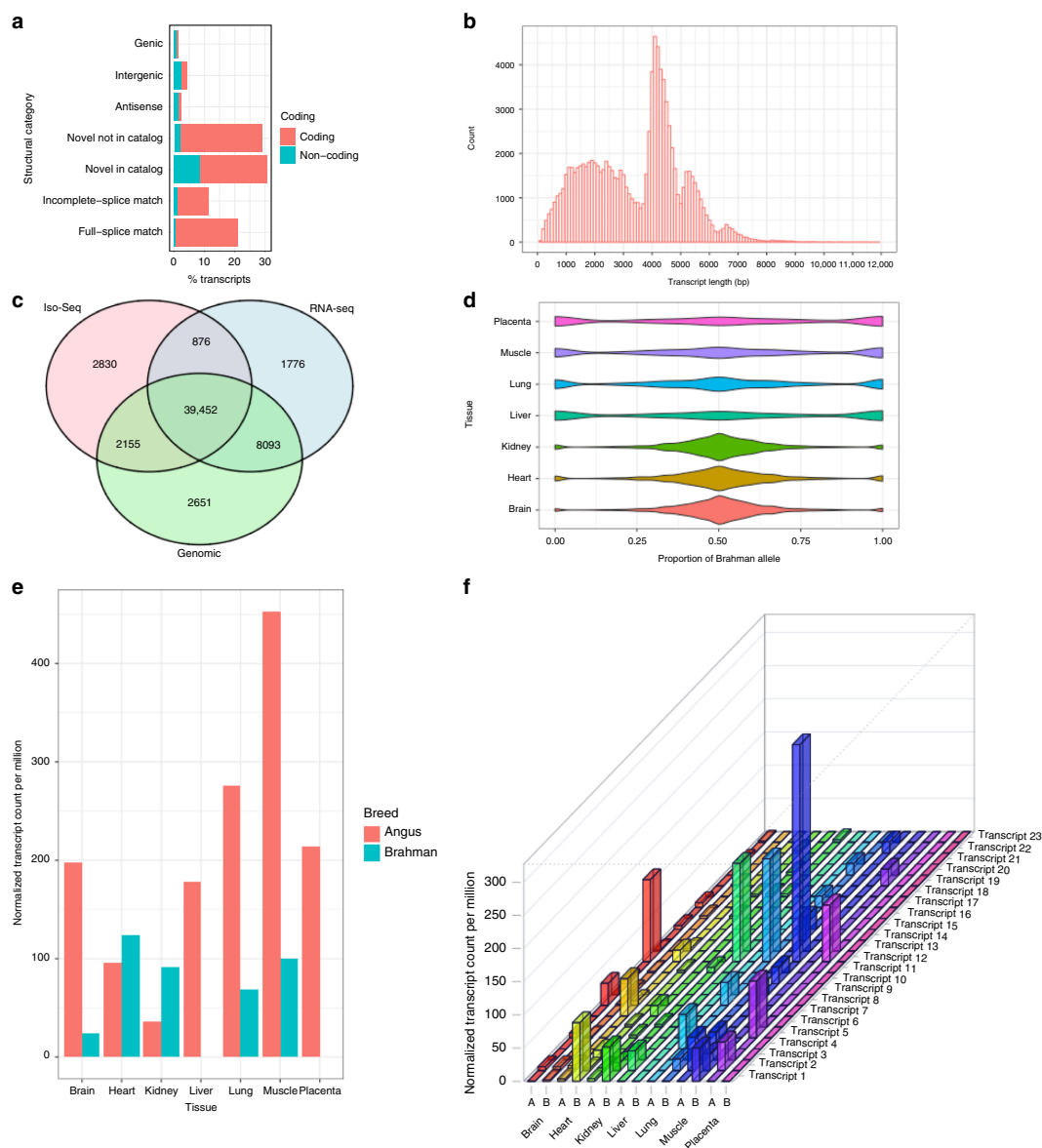


Fig. 6 Phasing of Iso-Seq full-length transcripts in seven tissues reveals transcriptional complexity and allelic imbalance. **a** Characterization of transcript annotation of the hybrid animal using SQANTI2 against the Brahman annotation. Full-splice match: perfect match with a reference; incomplete-splice match: missing one or more 5' exons against a reference; novel in catalog: novel combinations of known junctions; novel not in catalog: at least one novel splice site. **b** Histogram of transcript length distribution. **c** The overlap of SNPs between WGS short reads from genomic DNA, Iso-Seq, and RNA-Seq when Brahman was used as the reference genome. **d** Violin plot of the proportion of Brahman alleles, which was calculated as the normalized count of Brahman alleles divided by the sum of the normalized count of both Brahman and Angus alleles. Transcripts showing allelic imbalance and with higher expression in Brahman have values closer to 1, whereas those with higher expression in Angus have values closer to 0. **e** Tissue-specific allelic expression at the gene level for *ARIH2*, which is the most highly expressed Angus gene in the brain. **f** Tissue-specific allelic expression at the transcript level for *ARIH2* in the brain, heart, kidney, liver, lung, muscle and placenta. A denotes Angus and B denotes Brahman.

will facilitate the exploration of breed-specific effects on parental imprinting, which has been shown in maize¹⁸.

We found that the choice of reference assembly had a large impact on SV calling. The observed SV difference between

Brahman and Angus is in part due to using Hereford as the reference, which is more closely related to Angus. Ambiguous read alignments, which result from the assignment of reads to incorrect positions on the genome, are a major factor in SV call

accuracy¹⁹. This is a major concern in the detection of variant sites from alignment of reads to fasta-based reference assemblies. This has prompted the creation of graph-based file formats to improve alignment accuracy^{20,21}. After converting SVs from each assembly onto the Hereford assembly coordinates and calculating the intersection, we identified 1.3 Mbp of SVs (33% in genic regions) present in the Angus and Brahman assemblies that were not present in the Hereford assembly. This suggests that either the Hereford assembly was not as representative of the true structural variation in these regions or that there were assembly errors in the Angus and Brahman assembly that generated false positive SVs. The latter is less likely given the high accuracy of the Angus and Brahman genomes. V_{st} estimates for copy number windows incorrectly determined heightened variance between taurine and indicine animals on chromosome 15, suggesting that comparative alignment approaches are prone to a high FDR when used to detect true structural differences between species or subspecies of cattle. If only one reference genome is available for a genus or species, this could present a substantial issue in the interpretation of comparative SV analysis. Conversely, we identified 0.9 Mbp SVs shared between only the Hereford and Angus assembly, which may represent true genomic structural differences between taurine and indicine cattle.

HQ haplotype-specific assemblies facilitate genome-wide comparisons to identify novel variation. The discovery of an indicine-specific, additional copy of *FADS2P1*, which has been under positive selection, is an example that highlights the benefits of HQ haplotype-specific assemblies. The *FADS2P1* gene region in both Brahman and Angus span ~1.4 Mb of sequence, while the two *FADS2P1* genes in the water buffalo span ~1 Mb. The orthologous region in goat is ~1 Mb but contains gaps. Taking phylogenetic and information on conservation of synteny together, the most parsimonious explanation is that the extra *FADS2P1* was duplicated in the indicine lineage after divergence from taurine cattle. Rapid evolution at the *FADS2P1* locus resulted in neofunctionalization of the additional gene in indicine animals, with profound changes seen in the small exon 7.

FADS2 is a pleiotropic gene with known functions in the biosynthesis of unsaturated fatty acids, lipid homeostasis, inflammatory response, and promotion of myocyte growth and cell signaling^{22–24}. A non-synonymous SNP in exon 7 of Japanese Black cattle is significantly associated with linoleic acid²⁵ composition. While we do not know the functional significance of positively selected residues in the additional *FADS2P1* copy in Brahman, the SNP reported in the Japanese Black shows the importance of exon 7 in *FADS2* function. Studies in rats have shown that linoleic acid is an important component of skin ceramides and its deficiency increases water permeability of the skin²⁶. Comparisons between indicine and taurine animals have shown differences in fatty acids²⁷ and types of phosphatidylcholines²⁸. We hypothesize *Bos indicus* has three copies of *FADS2P1* genes to regulate the composition of fatty acids that constitute the cell membranes and could alter water permeability and heat loss from skin.

Brahman cattle may be better adapted to harsher environments because they have slower protein turnover²⁹. Relative to Angus, Brahman have much lower expression of *ARIH2* in key metabolic organs, such as the skeletal muscle, and no detectable expression in the liver. *ARIH2* promotes ubiquitylation of DCNLI, which is a co-E3 ligase that performs cullin neddylation, a process that regulates one-fifth of ubiquitin-dependent protein turnover¹⁴. CNV analysis revealed a decreased number of ubiquitin-conjugating enzyme E2D3 genes in the indicine lineage, which suggests lower protein turnover in indicine animals. While it is still speculative, our findings are consistent with lower protein turnover and the ability of Brahman cattle to withstand stressful conditions.

The analyses of CNV by alignment of short-read sequences from 38 individuals from 7 breeds to the Brahman and Angus genomes revealed that 6 genes with immune-related functions and putative roles in response to disease challenge and external parasites have additional copies in the indicine lineage. Conversely, *KRTAP9-2*, a gene with significantly altered gene expression following tick infestation³⁰, is expanded in the taurine lineage, which has also been reported in previous CNV studies^{13,31}. Further studies are needed to elucidate how changes in copy number of *KRTAP9-2* affect its expression and its role in tick resistance.

In conclusion, the approach used here is able to create haplotype-resolved genome assemblies that are of higher quality than traditional haplotype-collapsed assemblies. Availability of these HQ assemblies has enabled us to better resolve SVs and identify regions under selection that may be involved in adaptation to the environment. Looking forward, it is clear that HQ haplotype-resolved assemblies together with long-read transcript information will underpin studies on genome function, regulation, and the control of phenotypes.

Methods

***Bos taurus* hybrid.** A *Bos taurus indicus* female (Brahman) was inseminated with semen from a *Bos taurus taurus* (Angus) bull. The *indicus* maternal genetic background of the Brahman dam was confirmed by mitochondrial DNA haplotype analysis³². At day 153 post-insemination, dam and conceptus were ethically sacrificed and fetal brain, heart muscle, kidney, liver, lung, skeletal muscle, and placenta (cotyledon) tissue were snap frozen in liquid nitrogen and stored at -80°C until further use. All animal work was approved by the Animal Ethics Committee of the University of Adelaide (No. S-094-2005).

Genome sequencing and assembly of contigs. DNA was extracted from fetal lung using a salting out method⁶. Briefly, 100 mg of tissue was ground into powder under liquid nitrogen and then transferred to a tube containing nuclei lysis solution (2 ml buffer of 10 mM Tris-HCl pH 8, 0.4 M NaCl, 2 mM EDTA, 0.2 ml 10% sodium dodecyl sulfate, 0.06 ml 10 mg/ml RNase A). After mixing at 37°C for 1 h, 0.025 ml of Proteinase K (20 mg/ml) was added to the solution and shaken overnight, and DNA was precipitated by salting out. The dam uterus and bull semen DNA were extracted using standard phenol–chloroform procedures. Twelve SMRT sequencing libraries were made from the fetal DNA using the protocol recommended by the Pacific Biosciences (Procedure P/N 100-286-000-07), with a 15-kb size selection cut-off on a Blue Pippin instrument (Sage Science, Beverly, MA). Nine libraries were sequenced using P6/C4 chemistry on an RSII machine, whereas the remaining three libraries were sequenced on a Sequel machine. Approximately 161 Gb of RSII data and 205 Gb of Sequel data were produced, which gave a total sequence yield of 366 Gb with the mean read length of ~10.4 kb. Assuming a genome size of 2.7 Gb, the raw PacBio data represents ~136 \times coverage.

Illumina sequencing libraries for both parents (i.e., sire and dam) and F1 fetus were prepared using TruSeq PCR-free preparation kits (Illumina, San Diego, CA). A total of ~55 \times , ~60 \times , and ~84 \times coverage of 150 bp paired-end reads were generated for the sire, dam, and F1 fetus, respectively. In order to assemble phased haplotigs for the F1 Brahman–Angus hybrid, we used the trio binning method introduced by Koren et al.⁶. Briefly, 21-mers were identified in both sire and dam Illumina reads and 21-mers unique to one or other parent were used to assign the F1 PacBio long reads to the parent of origin. Approximately 1% of the PacBio reads were excluded from the assembly as they lacked parent-of-origin-specific 21-mers, due to their shorter lengths (Supplementary Fig. 9). Long reads that were binned into paternal and maternal groups were assembled separately with TrioCanu v1.6.

Hi-C library preparation and sequencing. A Sau3AI Hi-C library was prepared (Phase Genomics, Seattle, WA) as follows: approximately 200 mg of fetal lung tissue was finely chopped and then cross-linked in Proximo crosslinking solution. The 5' overhangs after Sau3AI digestion were filled with biotinylated nucleotides, and free blunt ends were ligated. After ligation, crosslinks were reversed and the free DNA was column purified and sonicated to approximately 600 bp peak fragment size (Bioruptor, Diagenode). Hi-C junctions were bound to streptavidin beads and washed to remove unbound DNA. Washed beads were used to prepare sequencing libraries using the HyperPrep Kit (Kapa) following the manufacturer's protocols. In total, 203 million 2 \times 81 bp read pairs were sequenced on NextSeq Illumina platform.

Scaffolding of contigs with Hi-C. All Hi-C reads were mapped to each breed-specific set of haplotigs using BWA v0.7.15³³. A haplotype score for a pair was defined as the sum of the percent identity multiplied by match length for each read end (unmapped read ends were assigned a score of 0). Each read pair had two scores, one

per haplotype. Pairs with a higher score for one haplotype were considered breed specific and assigned to their respective haplotype. Pairs with a tied score were considered homozygous and assigned to both haplotypes for scaffolding.

Three different Hi-C based scaffolding programs, 3D-DNA³⁴, Proximo (Phase Genomics), and SALSA2³⁵, were evaluated for scaffolding contigs. Further detail on the comparison between the scaffolders is given in Supplementary Note 1. Reads were mapped with the Arima mapping pipeline (https://github.com/ArmaGenomics/mapping_pipelinecommit72c81901c671203a86ca4675457004a71d0cd249) and converted to bed format prior to SALSA2 scaffolding (<https://github.com/machinegun/SALSAgitcommit863203dd094af9b342c35feede7dabec37b44>), which was run with parameters `-c 10000 -e GATC -m yes`, for both breed-specific haplotypes.

Bionano DNA isolation and assembly. DNA was extracted from 10 mg kidney tissue from the F1 hybrid using the Bionano Animal Tissue DNA Isolation Kit (P/N 80002) with slight modifications as follows: the frozen tissue was crushed in liquid nitrogen, placed in 2% formaldehyde in Bionano animal tissue homogenization buffer (Document number 30077, Bionano-Prep-Animal-Tissue-DNA-Isolation-Soft-Tissue-Protocol.pdf), and blended with a rotor-stator. The homogenate was passed through a 100- μ m nylon filter, fixed on ice for 30 min in 2 ml 100% ethanol, and centrifuged for 5 min at 2000 \times g. The resulting pellet was resuspended in homogenization buffer and added to pre-warmed agarose to make 0.8% agarose plugs. High molecular weight DNA was extracted from the agarose plugs, labeled, stained, and imaged on a Bionano Saphyr system³⁶. Further detail on de novo optical map assembly is given in Supplementary Note 2.

RNA-Seq and Iso-Seq. RNA was extracted from tissue and ground to a fine powder under liquid nitrogen using the Qiagen RNeasy Plus Universal Kit as per the manufacturer's instructions. RNA quality was assessed using an Agilent TapeStation system and confirmed as RNA integrity number >8 for all samples. Sequencing libraries were prepared with the KAPA Stranded RNA-Seq Library Preparation Kit as per the manufacturer's protocol and sequenced on an Illumina Next-Seq machine for 100 bp paired-end reads with the target of 50 million reads per sample.

Iso-Seq data were generated from brain, heart muscle, kidney, liver, lung, skeletal muscle, and placenta (cotyledon) tissue. Iso-Seq SMRT bell libraries were created according to the PacBio protocols. Briefly, two size-selected cDNA pools were created, one with an average cDNA size ~3 kb and the second with a cDNA size of ~7 kb. The two pools were then combined for SMRTbell™ Template Preparation. The final average library size was ~5 kb as measured by a bioanalyzer. Each SMRTbell library was loaded onto the Sequel at approximately 50 pM.

Identification and phasing of full-length transcripts. The Iso-Seq data was processed using the isoseq3.1.0 software on the PacBio Bioconda (<https://github.com/PacificBiosciences/pbioconda>). The process consists of (1) generating CCS reads, (2) classifying FLNC reads that have the 5', 3' cDNA primer sequence, and the polyA tail, (3) clustering FLNC reads at the isoform level and generating a draft consensus for each isoform, and (4) polishing each isoform to create HQ, full-length transcript sequences.

The HQ transcript sequences were then mapped to the Brahman reference genome using minimap2 (v2.15-r905) and filtered for alignments that had $\geq 99\%$ coverage and $\geq 95\%$ identity. Redundant and degraded transcripts were collapsed using the Cupcake tool (https://github.com/Magdoll/cDNA_Cupcake). SQANTI2³⁷ was used to annotate transcripts that belong to seven distinct categories: (i) known isoforms with full-splice match, (ii) known isoforms with incomplete-splice match, (iii) novel isoforms in catalog, (iv) novel isoforms not in catalog, (v) antisense transcripts, (vi) transcripts that overlap with intergenic region, (vii) transcripts that overlap with genic regions.

In order to phase transcripts using the Iso-Seq data, we ran IsoPhase, which is a part of the Cupcake tool, against the Brahman reference. IsoPhase first piles up the FLNC reads of all the isoforms of a gene and calls substitution SNPs using a one-sided Fisher exact test with Bonferroni correction at a P value cut-off of 0.01. It then infers haplotypes based on the phasing information provided by the FLNC reads. The output defines the inferred haplotypes for each transcript and estimates the relative abundance of each allele. We ran IsoPhase using the pooled set of all FLNC reads from all tissues, then later demultiplex them to create an abundance matrix that is specific for each haplotype, per isoform FLNC count for each tissue. To compare the abundance of transcripts across tissues, we normalized the counts by dividing the FLNC counts for each haplotype isoform by the total number of FLNC counts in that tissue, multiplied by a million to obtain the transcript per million number.

Mapping RNA-Seq and WGS reads from the F1 hybrid tissues. For RNA-Seq, read mapping was performed with Hisat2 v2.1.0³⁸, whereas the genomic short reads were mapped using BWA v0.7.15³³. SNPs were called using GATK v4³⁹.

Scaffold validation with recombination map. Scaffold contiguity was assessed using a previously published recombination map⁴⁰. Briefly, the recombination map probe sequences were aligned using BWA MEM v0.7.15 to the scaffolds and the coordinates were arranged in a directed acyclic graph, using a custom script. A contiguity break between consecutive recombination map-ordered probes in the scaffolds was considered an error; however, we tolerated one mismatched probe in a

window of three consecutive probes (Hamming distance = 1) to avoid false positive detection due to mapping ambiguity. Despite having Hi-C sequences, some scaffolds that belonged to chromosomes could not be joined together, which necessitated the use of recombination map markers to join and orientate these scaffolds.

Gap filling and polishing. After checking scaffolds with recombination maps⁴⁰, the Angus and Brahman scaffolds that contained 343 and 369 gaps, respectively, were gap filled with PBjelly⁴¹ v15.8.24 using haplotype-specific PacBio subreads. The default parameters of PBjelly were used, except for the support mode, where the options `captureOnly` and `spanOnly` were used. This step closed 52 and 61 gaps in Angus and Brahman scaffolds, respectively. Two rounds of ArrowGrid (<https://github.com/skoren/ArrowGrid>) was run to polish the scaffolds to give quality scores.

Assembly evaluation and genome annotation. The assemblies were evaluated with BUSCO v2.0.1⁴² and other metrics that include compression/expansion errors. Annotations were created using the Ensembl gene annotation system⁴³ and the NCBI pipeline. Further detail on the annotation process is given in Supplementary Notes 3 and 4, and for assembly evaluation, detail is given in Supplementary Note 5.

Repeat analysis. RepeatMasker version open-4.0.7 (<http://www.repeatmasker.org>) was used to search for repeats in the UOA_Angus_1 and UOA_Brahman_1 assemblies by identifying matches to RepBase (version RepBase23.10.embl)⁴⁴. Repeats in the current water buffalo assembly (UOA_WB_1) and cattle assembly (UMD3.1.1) were downloaded from the NCBI. Repeats with matches $\leq 60\%$ identity were filtered out. Centromeric repeats were identified by searching repeats that belonged to the family Satellite/centr in Repbase. We scored a sequence of repeat units as one block and counted the blocks, applying this method systematically throughout for all scaffolds. The vertebrate telomeric repeat, 6-mer TTAGGG, was identified by RepeatMasker. The search for at least 2 consecutive identical TTAGGG repeats within 1000 kb of chromosome ends was done to detect the presence of telomeres.

Gap comparisons and sequence contiguity. To evaluate gaps and sequence contiguity, the Angus and Brahman assemblies were compared to the water buffalo, human, and Hereford cattle assemblies. Only sequences that belong to autosomes and sex chromosomes were retained for analysis, whereas unplaced and mitochondrial sequences were filtered out. The tool `seqtk v1.2-r94` (<https://github.com/lh3/seqtk>) was used to count gaps with similar code implementation as those used for the water buffalo genome⁹.

SNP and indel calls. Thirty-eight individuals with ~10 \times WGS short-read Illumina data representing 7 breeds were selected from the USMARC Beef Diversity Panel version 2.9 (MBCDPv2.9)⁴⁵. The individuals selected for the panel were bulls with minimal pedigree relationships to maximize sampling of diverse alleles suitable for population genetics studies. The number of individuals per breed was as follows: six Angus, five Brahman, six Gelbvieh, six Hereford, five Red Angus, five Shorthorn, and five Simmental. These six taurine breeds were chosen on the basis that they were unlikely to carry *B. indicus* genetics given their history.

WGS data quality of each individual was checked with FASTQC v0.11.4⁴⁶ and then trimmed with Trim Galore v0.4.2⁴⁷ to a minimum length of 110 bp per read and Phred score of 20. Potential adapters in the sequence reads were removed using AdapterRemoval v2.2.1⁴⁸. Following trimming, the reads were checked with FASTQC again to ensure that only HQ reads were retained. Reads were then mapped to both the Angus and Brahman assemblies separately using BWA v0.7.15³³ with the option `mem`. Samtools v1.8⁴⁹ was used to convert the resulting alignment to sorted bam format. Duplicate reads, which may be due to PCR artifacts, were marked with Picard v2.2.4⁵⁰ MarkDuplicates. The bam files from each individual animal were merged with GATK v4³⁹ MergeSamFiles function. Then the following series of GATK functions, AddOrReplaceReadGroups, HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs, were applied to the alignment files to generate a variant call file in VCF v4.2 format. SNPs were filtered with VariantFiltration function using the parameters `(QD < 2.0) || (FS > 60.0) || (MQ < 40.0) || (MQRankSum < -12.5) || (ReadPosRankSum < -8.0)`. Indels were filtered with VariantFiltration function using the parameters `(QD < 2.0) || (FS > 200.0) || (ReadPosRankSum < -20.0)`. Annovar tool⁵¹ version dated 2017-07-17 was used to annotate the variants.

SV and copy number variant analyses. WGS short-read data sets from the same 38 animals used for SNP and indel calls were aligned to the UOA_Angus_1, UOA_Brahman_1, and ARS-UCD1.2 with BWA MEM v0.7.15³³ and further processed with Samtools v1.9⁴⁹. Read-pair and split-read profile SVs were called with the lumpy-sv v0.2.13⁵² pipeline, lumpyexpress, using default parameters for each sample. lumpy-sv VCF files were converted to BEDPE format using the `vcfToBedepe` script included in the lumpy-sv software package. Copy number estimates for genomic segments were calculated from normalized WGS read depth using JaRMS v0.0.13 as previously described⁵³. As JaRMS estimates of genomic copy number are distributed around a value of 1, as the normal diploid copy

number count, we multiplied the level estimates from the JaRMS program by two to obtain the adjusted copy number state of genomic regions. JaRMS copy number estimates were used to estimate the population differentiation of taurine and indicine cattle on a per-gene basis using the V_{st} metric^{12,13}. A custom script (CalculateVstDifferences.py) was used to automate the calculation of V_{st} and generation of data tables for plotting. Genes that had a $V_{st} > 0.3$, which is equivalent to the top 1% V_{st} , and a difference in average copy number between groups > 3 were considered to have a significant difference in copy number between taurine and indicine populations.

In addition to using short WGS reads from the 38 individuals of 7 breeds to find SVs, the haplotype-resolved Angus and Brahman genomes were aligned with the HQ ARS-UCD1.2 cattle reference to assess SVs. The advantage of aligning to ARS-UCD1.2 was to standardize the SVs specific to each haplotype on a common coordinate system. Contigs obtained by breaking final scaffolds at gap positions from UOA_Angus_1 and UOA_Brahman_1 were aligned using nucmer v4⁵⁴ to the ARS-UCD1.2 assembly to identify the larger structural differences (50–10,000 bp) using Assemblytics¹¹. The nucmer alignment parameters were `-maxmatch -t 4 -l 100 -c 500`, which was followed by delta-filter with the option `-g`. Assemblytics parameters followed the default settings, which were “Unique sequence length required: 10000, Maximum variant size: 10000, Minimum variant size: 50.” The overlap of SVs with Ensembl annotation of Hereford cattle ARS-UCD1.2 release 96 were identified with GenomicFeatures and systemPipeR R packages.

Identification, copy number and phylogenetic tree of FADS2P1. All chromosomes from Brahman were aligned to the corresponding Angus chromosomes using the dot plot tool Gepard v1.4⁵⁵. Genomic regions that differed between the two subspecies were isolated for further scrutiny. Of all the regions analyzed, one particular locus on Brahman chromosome 15 at position ~4 Mb covering ~200 kb diverged from the corresponding Angus chromosome. Further analysis revealed an extra copy of *FADS2P1* in the Brahman genome. BLASTP⁵⁶ analysis identified two copies of *FADS2P1* in Angus, Hereford, water buffalo, and goat, whereas only Brahman had three copies of this gene. A maximum likelihood tree with 1000 bootstraps was constructed for *FADS2P1* homologs using RAXML v8⁵⁷ with substitution model PROT-GAMMA-AUTO. The conservation of synteny around the *FADS2P1* locus was investigated by alignments of Angus to Brahman and Angus to Hereford using nucmer v4⁵⁴ and displayed with Ribbon⁵⁸.

Positive selection analysis on FADS2P1. The observation of an indicine-specific *FADS2P1* residing in a divergent region prompted further investigation into the possibility that the gene is under positive selection. Homologs of *FADS2P1* in Brahman, Angus, and water buffalo were subjected to CODEML analysis as implemented in PAML v4.8⁵⁹. Selective pressure acting on a gene can be estimated by the rate ratio (ω) of non-synonymous (amino acid changes) to synonymous (silent changes) substitutions. Detection of $\omega > 1$ is a sign of positive selection, and the site models, namely, M7 and M8 in PAML, which allow ω to vary among sites, were used to detect positive selection. Protein sequences of *FADS2P1* homologs were aligned using Muscle⁶⁰, and the corresponding nucleotides were mapped back onto the amino acid alignment using PAL2NAL⁶¹ with gap removal. The tree topology used to run CODEML was a maximum likelihood gene tree calculated from RAXML⁵⁷. Model M8 was compared with M7 using the likelihood ratio test (LRT) to evaluate if the model with positive selection was favored. More detail on similar positive selection methodology can be found in our study on mammalian glutathione S-transferases⁶².

Statistical analysis. R/Bioconductor was used for all statistical analyses. Significance of positively selected sites found in *FADS2P1* were evaluated using the LRT, with the test statistic $t_{LR} = 2[\ln(\text{Model 8}) - \ln(\text{Model 7})]$.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The PacBio reads, Hi-C reads, RNA-Seq, Iso-Seq, and Illumina paired-end reads are available in the SRA under BioProject PRJNA432857. The 38 individuals from seven breeds used for variant calls were downloaded from the BioProject PRJNA324822. The assemblies ARS-UCD1.2 (GCF_002263795.1), Bos_taurus_UMD_3.1.1 (GCF_000003055.6), ARS1 (GCF_001704415.1), and UOA_WB_1 (GCF_003121395.1) were downloaded from the NCBI. Intermediary assembly FASTA files and other miscellaneous information are available from the corresponding authors upon request. Annotation files of UOA_Angus_1 and UOA_Brahman_1 are available through Ensembl. Primary accession numbers: BioProject: PRJNA432857; GenBank assembly accession for UOA_Angus_1: GCA_003369685.2; and GenBank assembly accession for UOA_Brahman_1: GCA_003369695.2.

Code availability

Custom scripts can be found at GitHub repository at the following URL: <https://github.com/lloydlow/BrahmanAngusAssemblyScripts>.

Received: 20 August 2019; Accepted: 27 March 2020;

Published online: 29 April 2020

References

- Park, S. D. E. et al. Genome sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography and evolution of cattle. *Genome Biol.* **16**, 234 (2015).
- Verdugo, M. P. et al. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**, 173–176 (2019).
- Naik, S. N. Origin and domestication of Zebu cattle (*Bos indicus*). *J. Hum. Evol.* **7**, 23–30 (1978).
- Koufariotis, L. et al. Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. *Sci. Rep.* **8**, 17761 (2018).
- American Brahman Breeders Association. Available at <https://brahman.org> (2020).
- Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
- Cao, H. et al. De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).
- Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
- Low, W. Y. et al. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* **10**, 260 (2019).
- Zimin, A. V. et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10**, R42 (2009).
- Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
- Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
- Bickhart, D. M. et al. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA Res.* **23**, 253–262 (2016).
- Kelsall, I. R. et al. Coupled monoubiquitylation of the co-E3 ligase DCN1 by Ariadne-RBR E3 ubiquitin ligases promotes cullin-RING ligase complex remodeling. *J. Biol. Chem.* **294**, 2651–2664 (2019).
- Berchtold, M. W. & Villalobo, A. The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer. *Biochim. Biophys. Acta Mol. Cell Res.* **1843**, 398–435 (2014).
- Loftan, M. et al. Primary structures of different isoforms of buffalo pregnancy-associated glycoproteins (BuPAGs) during early pregnancy and elucidation of the 3-dimensional structure of the most abundant isoform BuPAG 7. *PLoS ONE* **13**, e0206143 (2018).
- Kim, J. et al. The genome landscape of indigenous African cattle. *Genome Biol.* **18**, 34 (2017).
- Wang, B. et al. Variant phasing and haplotypic expression from single-molecule long-read sequencing in maize. *Commun. Biol.* **3**, 1–11 (2020).
- Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
- Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
- Eggertsson, H. P. et al. GraphTyper enables population-scale genotyping using pan-genome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
- Gol, S. et al. polymorphism in the fatty acid desaturase-2 gene is associated with the arachidonic acid metabolism in pigs. *Sci. Rep.* **8**, 14336 (2018).
- Markworth, J. F. et al. Arachidonic acid supplementation modulates blood and skeletal muscle lipid profile with no effect on basal inflammation in resistance exercise trained men. *Prostaglandins Leukot. Essent. Fat. Acids* **128**, 74–86 (2018).
- Markworth, J. F. & Cameron-Smith, D. Arachidonic acid supplementation enhances in vitro skeletal muscle cell growth via a COX-2-dependent pathway. *Am. J. Physiol. Physiol.* **304**, C56–C67 (2013).
- Takahashi, H. et al. Association of bovine fatty acid desaturase 2 gene single-nucleotide polymorphisms with intramuscular fatty acid composition in Japanese Black steers. *Open J. Anim. Sci.* **06**, 105–115 (2016).
- Hansen, H. S. & Jensen, B. Essential function of linoleic acid esterified in acylglucosylceramide and acylceramide in maintaining the epidermal water permeability barrier. Evidence from feeding studies with oleate, linoleate, arachidonate, columbinatate and α -linolenate. *Biochim. Biophys. Acta Lipids Lipid Metab.* **834**, 357–363 (1985).
- Bressan, M. C. et al. Genotype x environment interactions for fatty acid profiles in *Bos indicus* and *Bos taurus* finished on pasture or grain. *J. Anim. Sci.* **89**, 221–232 (2011).
- Sudano, M. J. et al. Phosphatidylcholine and sphingomyelin profiles vary in *Bos taurus indicus* and *Bos taurus taurus* in vitro- and in vivo-produced blastocysts. *Biol. Reprod.* **87**, 130 (2012).

29. Sainz, R. D., Barioni, L. G., Paulino, P. V. R., S.C. Valadares & Filho, J. W. *Growth Patterns of Nelore vs. British Beef Cattle Breeds Assessed using a Dynamic, Mechanistic Model of Cattle Growth and Composition* (eds Kebreab, E., Dijkstra, J., Bannink, A., Gerrits, W. J. J. & France, J.) Ch. 16 (CAB eBooks, 2006).
30. Wang, Y. H. et al. Gene expression profiling of Hereford Shorthorn cattle following challenge with *Boophilus microplus* tick larvae. *Aust. J. Exp. Agric.* **47**, 1397 (2007).
31. Bickhart, D. M. et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* **22**, 778–90 (2012).
32. Hiendler, S., Lewalski, H. & Janke, A. Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. *Cytogenet. Genome Res.* **120**, 150–156 (2008).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
35. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
36. Formenti, G. et al. SMRT long reads and direct label and stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*). *Gigascience* **8**, (2019).
37. Tardaguila, M. et al. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. Preprint at <https://doi.org/10.1101/118083> (2017).
38. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
39. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
40. Ma, L. et al. Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet.* **11**, e1005387 (2015).
41. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
42. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–2 (2015).
43. Aken, B. L. et al. The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
44. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
45. Heaton, M. P. et al. Using diverse U.S. beef cattle genomes to identify missense mutations in EPAS1, a gene associated with high-altitude pulmonary hypertension. *F1000Research* **5**, 2003 (2016).
46. Andrews, S. FastQC: a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
47. Krueger, F. Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
48. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
49. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. Broad Institute. Picard tools. Broad Institute, GitHub repository. <http://broadinstitute.github.io/picard/> (2020).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
53. Oldeschulte, D. L. et al. Annotated draft genome assemblies for the Northern Bobwhite (*Colinus virginianus*) and the scaled quail (*Callipepla squamata*) reveal disparate estimates of modern genome diversity and historic effective population size. *G3 (Bethesda)* **7**, 3047–3058 (2017).
54. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
55. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
57. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
58. Nattestad, M., Chin, C.-S. & Schatz, M. C. Ribbon: visualizing complex genome alignments and structural variation. Preprint at <https://doi.org/10.1101/082123> (2016).
59. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
60. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–7 (2004).
61. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, 609–612 (2006).
62. Tan, H. M. & Low, W. Y. Rapid birth-death evolution and positive selection in detoxification-type glutathione S-transferases in mammals. *PLoS ONE* **13**, e0209336 (2018).

Acknowledgements

This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. The work was part funded by the JS Davies bequest to the University of Adelaide. We thank Bob Lee, Kristen Kuhn, Kelsey McClure, and William Thompson for technical assistance. We acknowledge funding from the Wellcome Trust (108749/Z/15/Z), the Biotechnology and Biological Sciences Research Council (BB/M011615/1 and BB/S020152/1) and the European Molecular Biology Laboratory. The work was supported in part by funds from USDA-ARS Project Number 3040-31320-012-00D. The use of trade names or commercial products in this manuscript is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity employer and provider. A.R., S.B.K., and A.M.P. were supported by the Intramural Research Program of the National Human Genome Research Institute, US National Institutes of Health. A.R. was also supported by the Korean Visiting Scientist Training Award (KVSTA) through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (HI17C2098). The work of F.T.-N. was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>).

Author contributions

J.L.W., T.P.L.S., A.M.P., and S.H. conceived and managed the project; T.P.L.S. generated long- and short-read genomic data, as well as RNA-Seq and Iso-Seq data; W.Y.L. analyzed all results; R.T. and C.L. validated sex chromosomes and provided guidance on gene expression work; S.B.K., A.R., D.M.B., B.D.R., Z.N.K., S.B.K., J.G., and M.P.H. were involved in genome assembly and scaffolding; A.R.H., J.L., and A.W.C.P. provided optical map data; D.M.B. performed CNV analysis; E.T. provided Iso-Seq FLNC and IsoPhase analysis; F.T.-N., F.J.M., and K.B. annotated the genomes; W.Y.L. and J.L.W. drafted the manuscript; and all authors read, edited, and approved the final manuscript.

Competing interests

S.B.K., Z.N.K., and E.T. are employees of Pacific Biosciences. A.R.H., J.L., and A.W.C.P. are employees of BioNano Genomics. J.G. is an employee of Dovetail Genomics. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-15848-y>.

Correspondence and requests for materials should be addressed to S.H., T.P.L.S. or J.L.W.

Peer review information *Nature Communications* thanks Fritz Sedlacek and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

6.1 Supplementary Figures

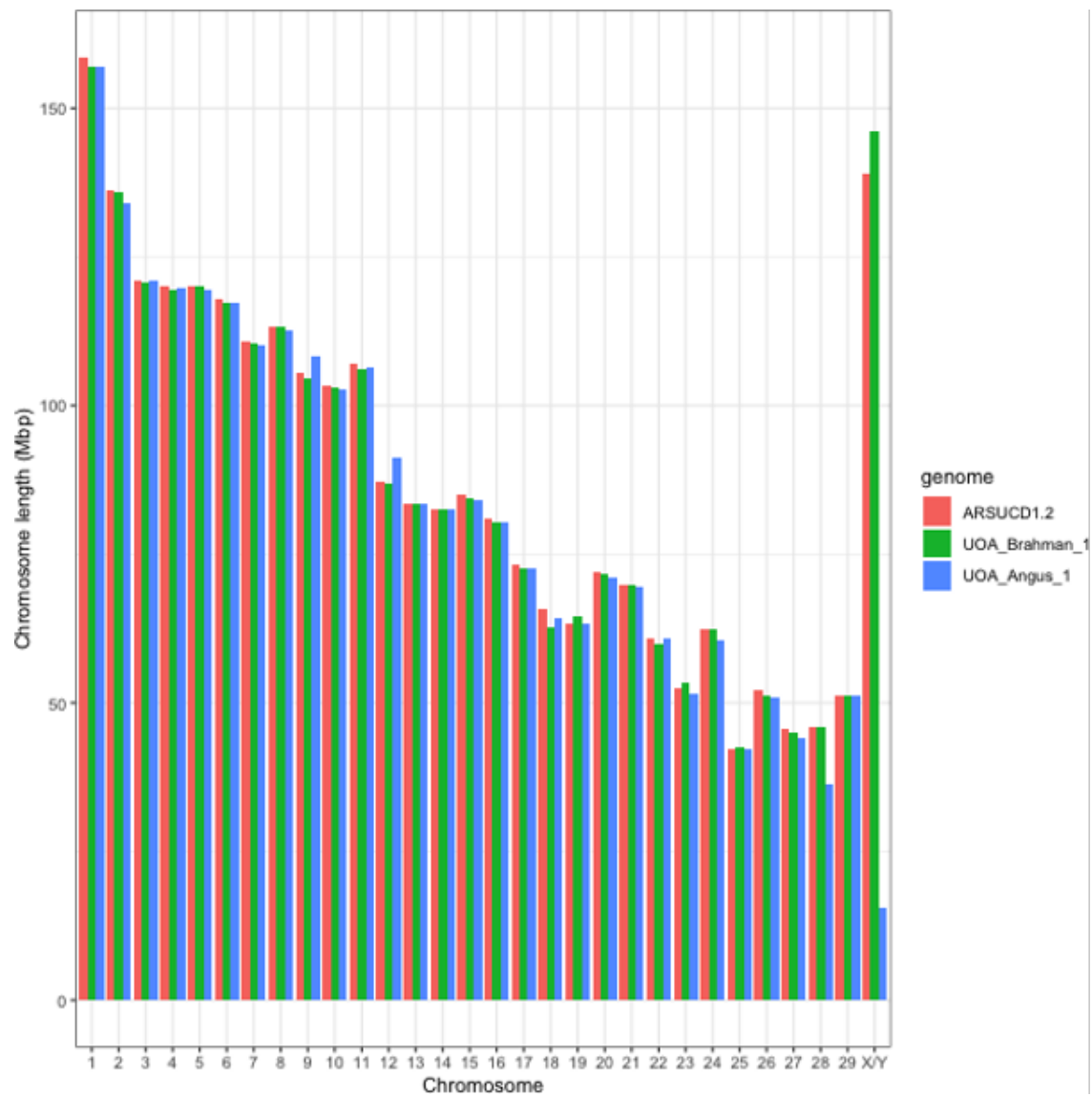


Figure 6.1: Comparison of chromosome sizes between Angus, Brahman and Hereford assemblies.

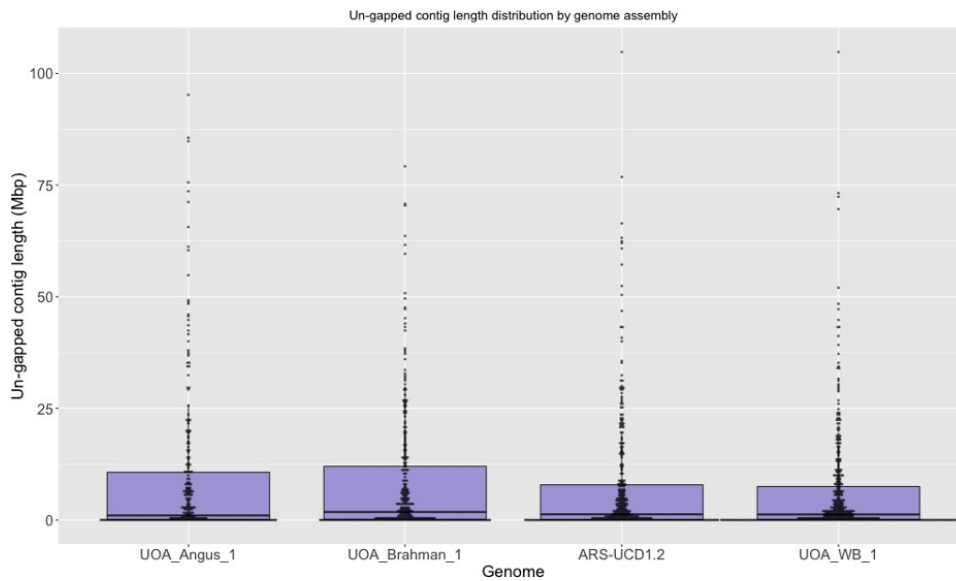


Figure 6.2: Distribution of un-gapped contig lengths in the three cattle breeds (UOA_Angus_1, UOA_Brahman_1 and ARS-UCD1.2) and water buffalo (UOA_WB_1).

Dot plots of individual values are overlaid on top of boxplots to show minima and maxima as circles. The bounds of box show the 25th and 75th percentile, with the median drawn as a thick line between these two quartiles.

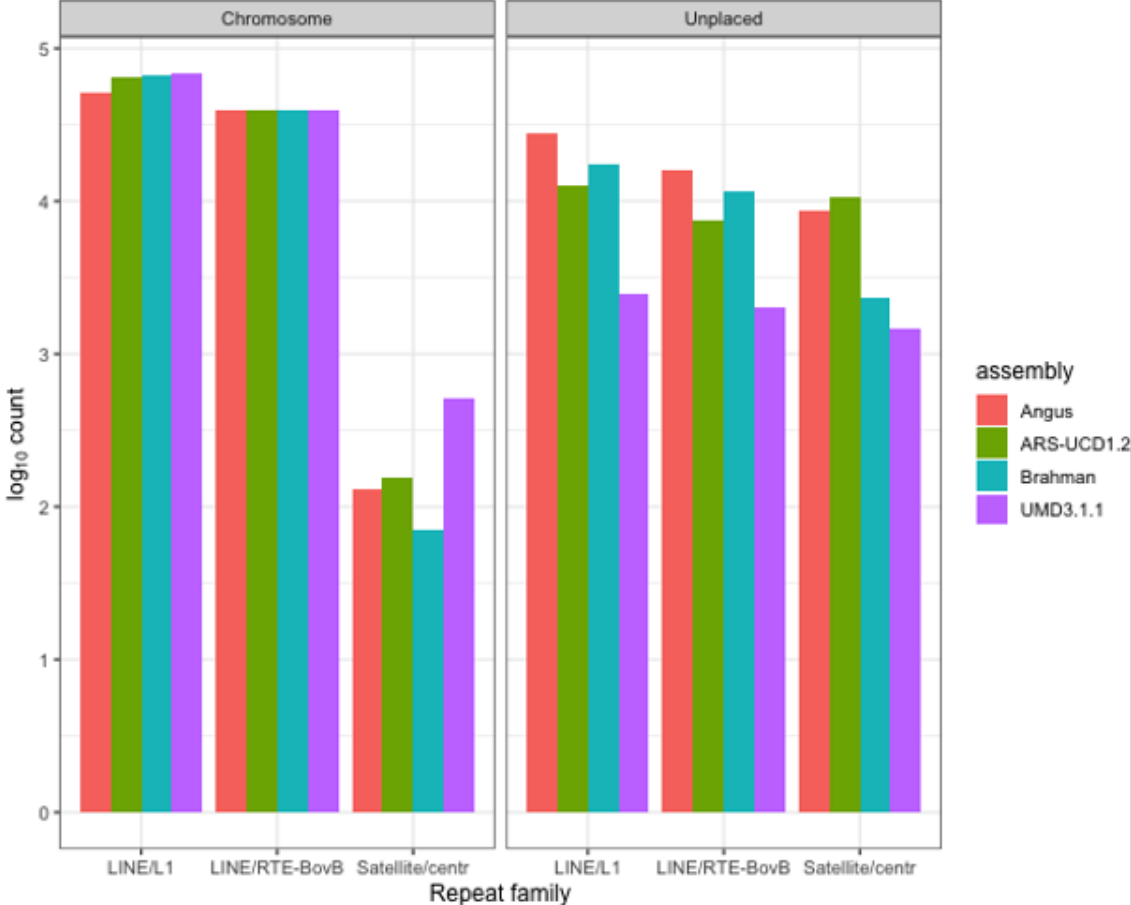


Figure 6.3: The count in (log₁₀ scale) of LINE/L1, LINE/RTE-BovB and Satellite/centromeric repeats in cattle genome assemblies.

The count for Satellite/centromeric is of a sequence of repeat units.

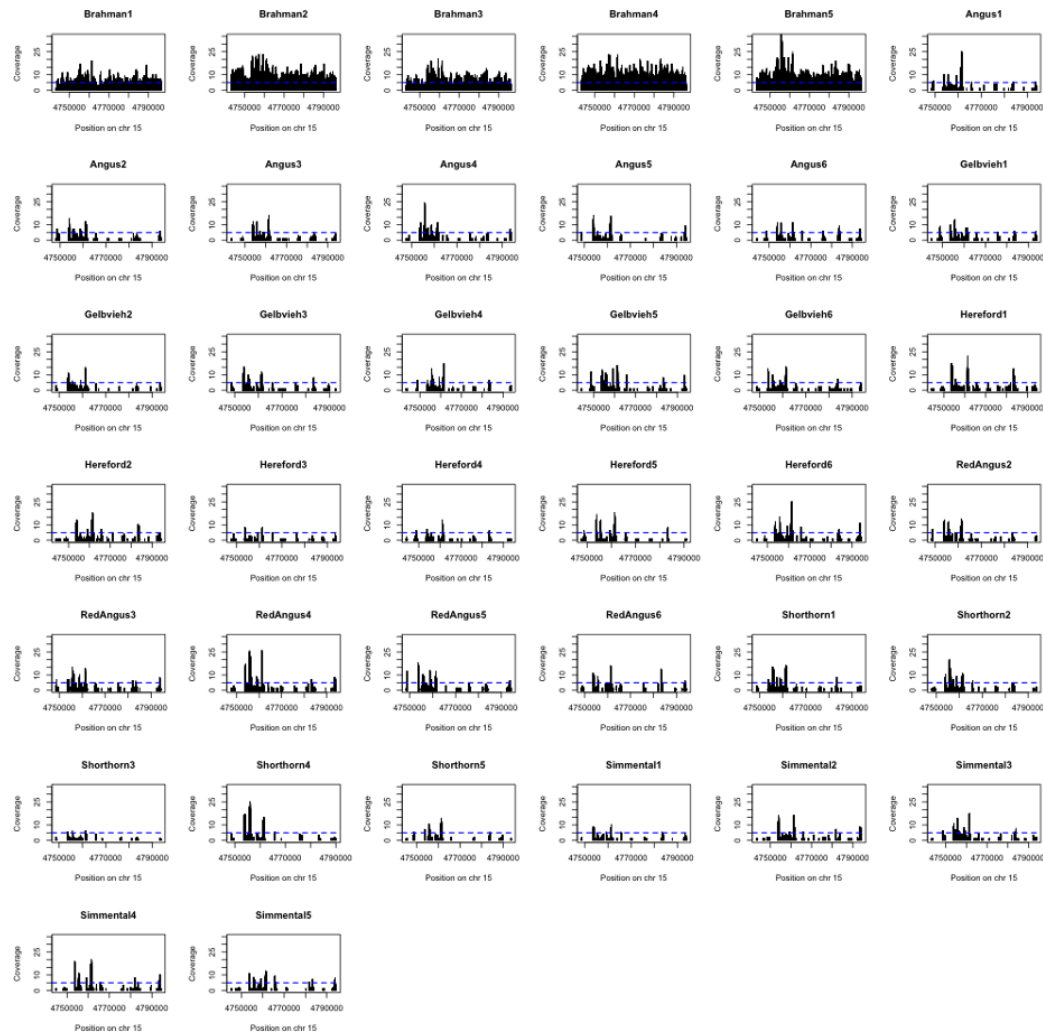


Figure 6.4: Coverage plot of *FADS2P1* in individuals of different cattle breeds.

The dashed blue line indicates the expected haploid coverage. As *FADS2P1* is member of a gene family, short reads that belong to other gene family members could potentially have mis-mapped to this region, which explains the non-zero coverage in taurine breeds at certain positions across the gene.

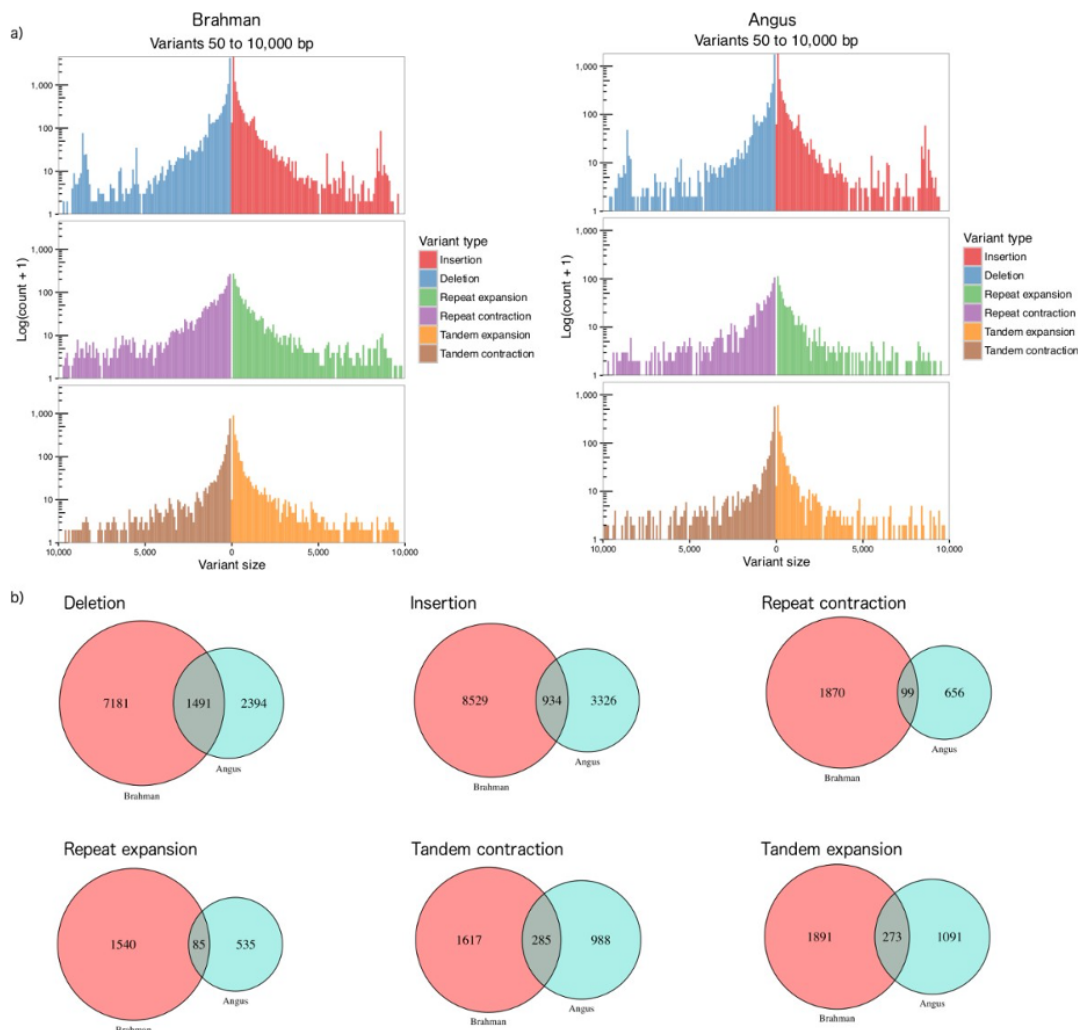


Figure 6.5: Distribution and breed specificity of Brahman and Angus structural variants.

a) Count of structural variants (SVs) categorized as deletion, insertion, repeat contraction, repeat expansion, tandem contraction, and tandem expansion by Assemblytics. The Hereford ARS-UCD1.2 was used as the common reference to call SVs in both Brahman and Angus contigs. b) Venn diagrams showing overlap of six classes of SVs between Brahman and Angus.

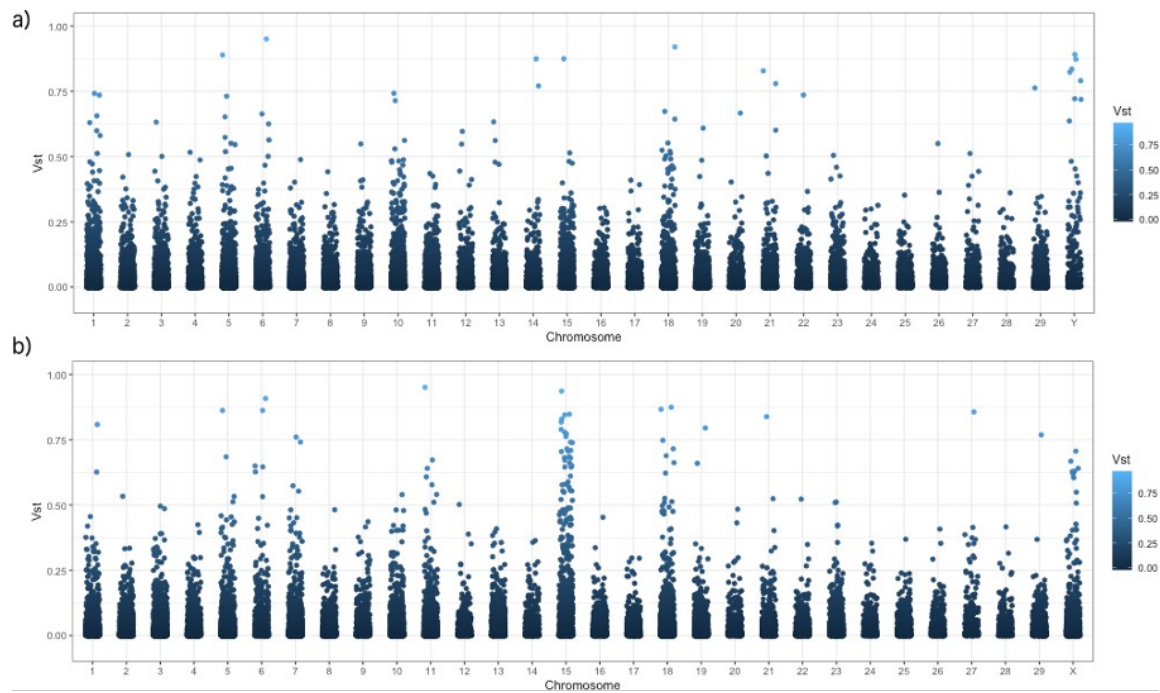


Figure 6.6: Analysis of copy number variations using different reference assemblies.

Population differentiation for copy number variations (CNV) as estimated by VST along each chromosome for the taurine and indicine comparison using a) UOA_Angus_1 and b) ARS-UCD1.2 as the reference genome.

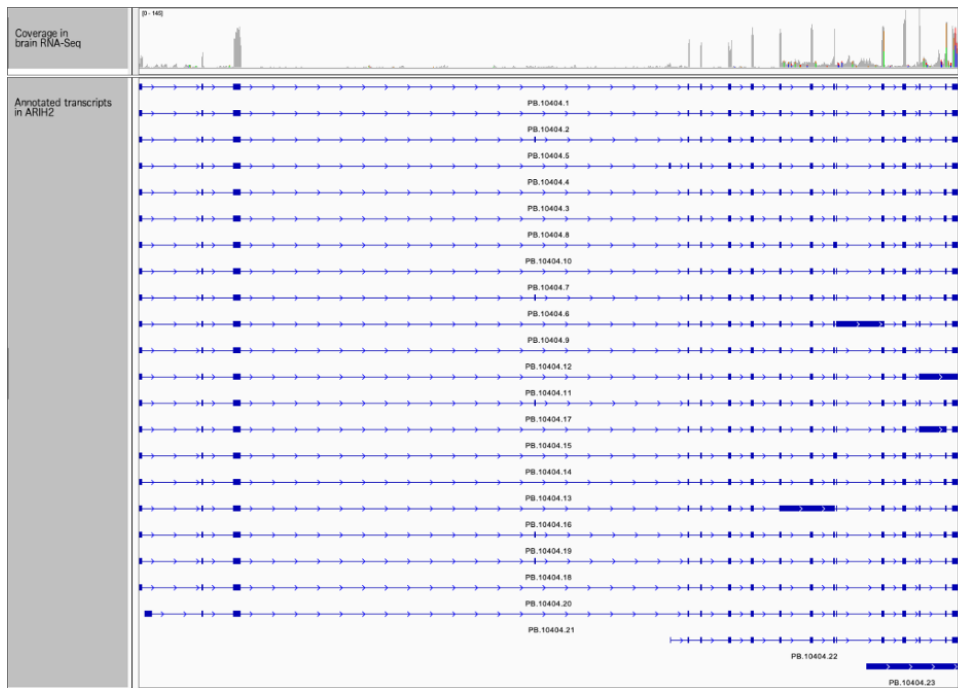


Figure 6.7: Full-length Iso-Seq transcripts (bottom) and RNA-Seq coverage for *ARIH2* in the brain tissue (top).

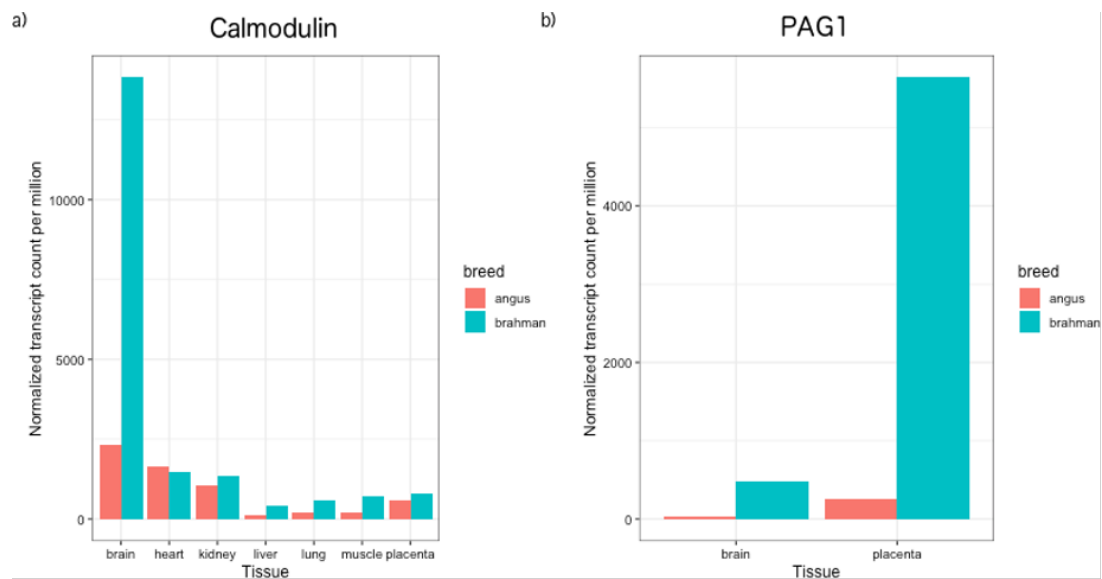


Figure 6.8: Normalized tissue-specific transcript counts for genes with allelic imbalance and higher expression of the Brahman allele in brain.

a) Calmodulin. b) Pregnancy-associated glycoprotein 1 (*PAG1*).

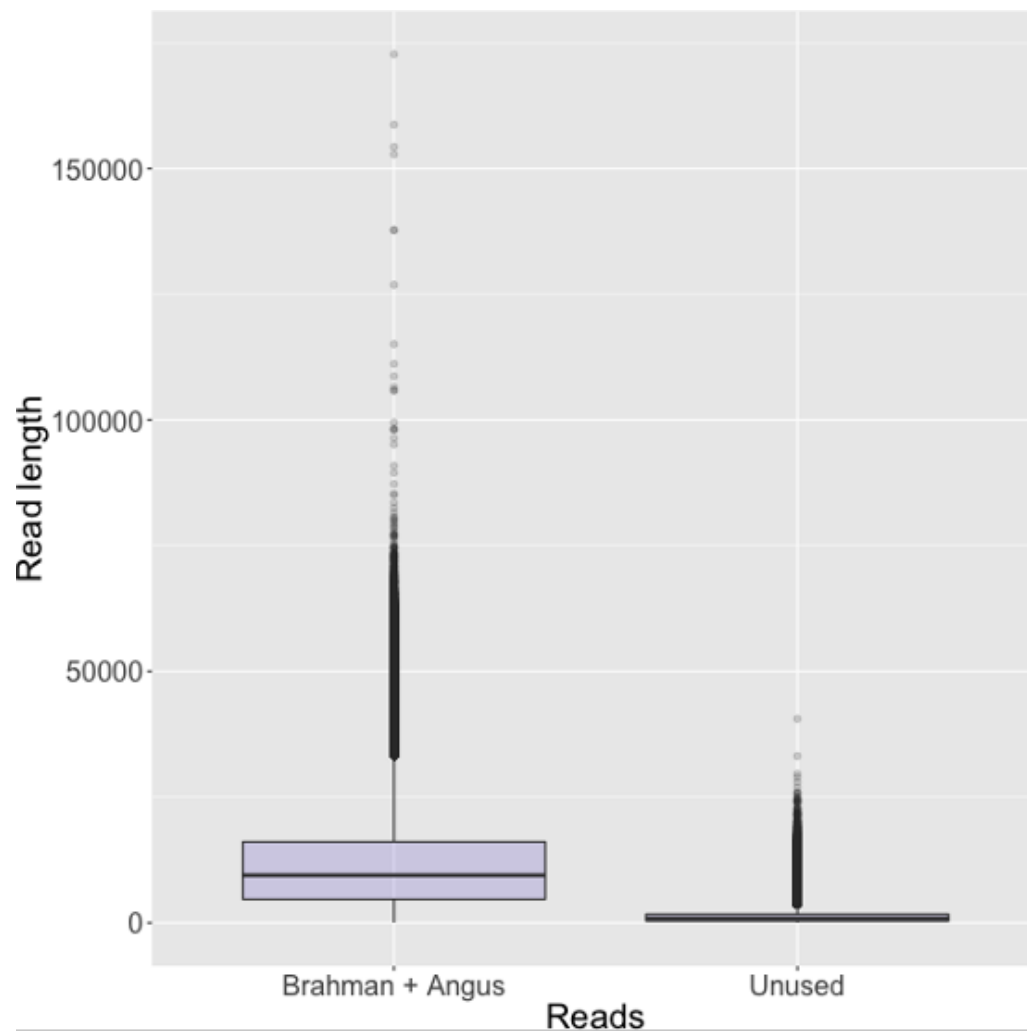


Figure 6.9: Distribution of unassigned PacBio WGS read length.

For each boxplot, the minimum is represented as the end point of the vertical line extending downward from the box whereas the maximum is the highest value shown as a circle. The bounds of box show the 25th and 75th percentile, with the median drawn as a thick line between these two quartiles.

6.2 Supplementary Tables

Table 6.1: Annotation features in Brahman, Angus and Hereford assemblies.

Feature	Brahman autosomes	Angus autosomes	Brahman X	Angus Y
gene	28547	28758	1363	192
lncRNA	3224	3269	153	25
miRNA	776	867	71	1
misc RNA	361	359	25	0
processed pseudogene	89	95	13	2
protein coding	21170	21266	948	153
pseudogene	458	451	38	3
rRNA	411	392	13	2
ribozyme	7	6	0	0
sRNA	3	3	0	0
scaRNA	32	31	1	0
snRNA	1130	1141	65	5
snoRNA	748	751	36	1

Comparison of UOA_Brahman_1, UOA_Angus_1, and ARS-UCD1.2 features annotated by the EMBL-EBI pipeline release 96. The columns with X and Y next to each breed show the annotated features in sex chromosomes of the corresponding assemblies.

Table 6.2: Assembly quality score values.

Statistic	Description	Angus	Brahman
QV	Quality value	44.63	46.38
COMPR_PE	Low CE-statistics computed on PE reads	211314	211783
STRECH_PE	High CE-statistics computed on MP reads	106768	92494
LOW_COV_PE	Low read coverage areas	61490	81218
LOW_NORM_COV_PE	Low paired-read coverage areas	58223	79034
HIGH_COV_PE	High read coverage areas	4329	3920
HIGH_NORM_COV_PE	High paired-read coverage areas	3803	2775
HIGH_SPAN_PE	High number of PE reads with pair mapped in a different scaffold	1808	2464
HIGH_SINGLE_PE	High number of PE reads with unmapped pair	30	96
HIGH_OUTIE_PE	High number of mis-oriented or too distant PE reads	15	10

Note: CE, compression/expansion; PE, paired-end.

Table 6.3: BUSCO assessment of the completeness of single-copy orthologs for Angus and Brahman genomes.

Description	Angus	Brahman
Complete BUSCOs	3813	3839
Complete and single-copy BUSCOs	3764	3790
Complete and duplicated BUSCOs	49	49
Fragmented BUSCOs	130	123
Missing BUSCOs	161	142
Total BUSCO groups searched	4104	4104
BUSCO completeness (%)	92.9	93.5

Table 6.4: Site models of CODEML for FADS2P1 and positively selected sites.

Model	Log-likelihood	$2\Delta(\ln L)$	P-value ^b	Positively selected sites ^a	Tree length	Average dN	Average dS
M7	-3149.50						
				237V			
				271V			
				294H			
				305C			
				306T			
				307V**			
				311L			
M8	-3132.38	17.12	0.00019	312F	0.5777	0.0168	0.0193
				315V			
				317L			
				324A			
				327C**			
				328R			
				329R			
				330S			
				370P			

^aBayes Empirical Bayes (BEB) was used to calculate posterior probabilities and only those with $\text{Prob}(\omega > 1) > 0.95$ are shown. ** indicates those with $\text{Prob}(\omega > 1) > 0.99$. Amino acid position follows Brahman ENSBIXP00005018486.1, which is the indicus-specific copy of *FADS2P1*. ^bP-value from likelihood ratio test.

Table 6.5: Genes identified in the selective sweep intervals.

Chr	Start	End	Ensembl ID	Name	biotype	Indicine	Taurine
						mean proportion alternate allele	mean proportion alternate allele
1	3500000	3600000	ENSBIXG00005005115	SCAF4	protein coding	0.048	0.401
1	81300000	81400000	ENSBIXG00005003979	SENP2	protein coding	0.093	0.405
1	81300000	81400000	ENSBIXG00005023240	LIPH	protein coding	0.093	0.405
1	81300000	81400000	ENSBIXG00005023174	RF00001	rRNA	0.093	0.405
1	107100000	107200000	ENSBIXG00005012541	not available	protein coding	0.1	0.584
1	136400000	136500000	ENSBIXG00005025046	ACAD11	protein coding	0.073	0.482
1	136400000	136500000	ENSBIXG00005024967	DNAJC13	protein coding	0.073	0.482
2	24300000	24400000	ENSBIXG00005005621	HAT1	protein coding	0.092	0.41
2	24300000	24400000	ENSBIXG00005027305	not available	protein coding	0.092	0.41
2	24300000	24400000	ENSBIXG00005027277	SLC25A12	protein coding	0.092	0.41
3	600000	700000	ENSBIXG00005024597	GPR161	protein coding	0.1	0.483
3	600000	700000	ENSBIXG00005024283	DCAF6	protein coding	0.1	0.483
3	600000	700000	ENSBIXG00005004103	RF00201	snoRNA	0.1	0.483
3	600000	700000	ENSBIXG00005023946	not available	protein coding	0.1	0.483
3	53800000	53900000	ENSBIXG00005030064	LRRC8C	protein coding	0.1	0.447
3	53800000	53900000	ENSBIXG00005030089	LRRC8B	protein coding	0.1	0.447
4	400000	500000	ENSBIXG00005018118	ESYT2	protein coding	0.092	0.563
4	400000	500000	ENSBIXG00005018040	NCAPG2	protein coding	0.092	0.563
4	1000000	1100000	ENSBIXG00005002902	not available	protein coding	0.064	0.459
4	52900000	53000000	ENSBIXG00005014470	not available	protein coding	0.043	0.409
4	72600000	72700000	ENSBIXG00005028158	CDHR3	protein coding	0.02	0.548
4	72600000	72700000	ENSBIXG00005004834	not available	miRNA	0.02	0.548
4	95100000	95200000	ENSBIXG00005018443	CRPPA	protein coding	0.083	0.474
5	42800000	42900000	ENSBIXG00005004767	FGD4	protein coding	0.07	0.537
5	42800000	42900000	ENSBIXG00005027380	RF00026	snRNA	0.07	0.537
5	48900000	49000000	ENSBIXG00005023898	SYN3	protein coding	0.091	0.591
5	106500000	106600000	ENSBIXG00005000567	not available	protein coding	0.1	0.518
6	116700000	116800000	ENSBIXG00005005681	not available	protein coding	0.07	0.409
6	116700000	116800000	ENSBIXG00005008386	not available	lncRNA	0.07	0.409
6	116700000	116800000	ENSBIXG00005008383	not available	protein coding	0.07	0.409
6	116700000	116800000	ENSBIXG00005000173	not available	lncRNA	0.07	0.409
6	116700000	116800000	ENSBIXG00005008365	FGFRL1	protein coding	0.07	0.409
6	116700000	116800000	ENSBIXG00005008347	not available	lncRNA	0.07	0.409
6	116700000	116800000	ENSBIXG00005008330	IDUA	protein coding	0.07	0.409
7	50000000	50100000	ENSBIXG00005019655	not available	lncRNA	0.04	0.437
7	50000000	50100000	ENSBIXG00005003205	IL17B	protein coding	0.04	0.437
7	50000000	50100000	ENSBIXG00005003199	PCYOX1L	protein coding	0.04	0.437

Continued on next page

Table 6.5 – continued from previous page

Chr	Start	End	Ensembl ID	Name	biotype	Indicine	Taurine
						mean proportion alternate allele	mean proportion alternate allele
7	50000000	50100000	ENSBIXG00005019586	not available	protein coding	0.04	0.437
7	50000000	50100000	ENSBIXG00005003178	AFAP1L1	protein coding	0.04	0.437
7	50000000	50100000	ENSBIXG00005019455	not available	protein coding	0.04	0.437
7	51900000	52000000	ENSBIXG00005003009	JAKMIP2	protein coding	0.1	0.405
7	66100000	66200000	ENSBIXG00005000684	HSPA4	protein coding	0.1	0.459
7	81100000	81200000	ENSBIXG00005020813	not available	lncRNA	0.085	0.47
8	7300000	7400000	ENSBIXG00005029063	DEFB136	protein coding	0.014	0.414
8	7300000	7400000	ENSBIXG00005029055	CTSB	protein coding	0.014	0.414
8	7300000	7400000	ENSBIXG00005029005	FDFT1	protein coding	0.014	0.414
8	10800000	10900000	ENSBIXG00005027580	not available	protein coding	0.1	0.431
8	10800000	10900000	ENSBIXG00005027560	ESCO2	protein coding	0.1	0.431
8	10800000	10900000	ENSBIXG00005004734	CCDC25	protein coding	0.1	0.431
8	52900000	53000000	ENSBIXG00005006324	VPS13A	protein coding	0.04	0.495
8	57500000	57600000	ENSBIXG00005011007	not available	protein coding	0.055	0.73
8	85000000	85100000	ENSBIXG00005000168	not available	pseudogene	0.053	0.433
9	32400000	32500000	ENSBIXG00005010729	CEP85L	protein coding	0.086	0.447
9	40600000	40700000	ENSBIXG00005008671	FIG4	protein coding	0.044	0.468
9	40600000	40700000	ENSBIXG00005008574	AK9	protein coding	0.044	0.468
9	85800000	85900000	ENSBIXG00005003851	SASH1	protein coding	0.083	0.475
10	31600000	31700000	ENSBIXG00005004621	not available	protein coding	0.092	0.501
10	31600000	31700000	ENSBIXG00005004584	CAPN3	protein coding	0.092	0.501
10	72600000	72700000	ENSBIXG00005007972	GPHN	protein coding	0.033	0.458
10	102100000	102200000	ENSBIXG00005017943	TTC7B	protein coding	0.1	0.478
10	102100000	102200000	ENSBIXG00005002910	RF00614	snoRNA	0.1	0.478
11	37900000	38000000	ENSBIXG00005017832	CFAP36	protein coding	0.1	0.498
11	37900000	38000000	ENSBIXG00005017770	PPP4R3B	protein coding	0.1	0.498
11	37900000	38000000	ENSBIXG00005017706	PNPT1	protein coding	0.1	0.498
11	45700000	45800000	ENSBIXG00005001920	NCK2	protein coding	0.075	0.406
11	45700000	45800000	ENSBIXG00005001916	not available	lncRNA	0.075	0.406
11	45700000	45800000	ENSBIXG00005014366	not available	lncRNA	0.075	0.406
11	45700000	45800000	ENSBIXG00005014363	RF00619	snRNA	0.075	0.406
11	45700000	45800000	ENSBIXG00005006127	not available	lncRNA	0.075	0.406
11	45800000	45900000	ENSBIXG00005007270	TTL	protein coding	0.1	0.438
13	59700000	59800000	ENSBIXG00005022377	PIP4K2A	protein coding	0.06	0.441
14	46600000	46700000	ENSBIXG00005016564	EXT1	protein coding	0.087	0.432
14	47000000	47100000	ENSBIXG00005007032	MED30	protein coding	0.022	0.461
14	64800000	64900000	ENSBIXG00005011378	VPS13B	protein coding	0.092	0.413
14	64800000	64900000	ENSBIXG00005011322	RF00156	snoRNA	0.092	0.413

Continued on next page

Table 6.5 – continued from previous page

Chr	Start	End	Ensembl ID	Name	biotype	Indicine	Taurine
						mean proportion alternate allele	mean proportion alternate allele
15	68600000	68700000	ENSBIXG00005002944	AMOTL1	protein coding	0.085	0.499
16	4300000	4400000	ENSBIXG00005010567	MAPKAPK2	protein coding	0.085	0.473
16	4300000	4400000	ENSBIXG00005010507	not available	protein coding	0.085	0.473
16	36100000	36200000	ENSBIXG00005023955	ATP1B1	protein coding	0.082	0.495
16	36100000	36200000	ENSBIXG00005023904	RF00155	snoRNA	0.082	0.495
16	36100000	36200000	ENSBIXG00005023880	NME7	protein coding	0.082	0.495
16	36100000	36200000	ENSBIXG00005004083	not available	protein coding	0.082	0.495
16	48600000	48700000	ENSBIXG00005016454	not available	protein coding	0.073	0.543
16	48600000	48700000	ENSBIXG00005002621	DFFB	protein coding	0.073	0.543
16	48600000	48700000	ENSBIXG00005016419	CEP104	protein coding	0.073	0.543
16	48600000	48700000	ENSBIXG00005002608	not available	miRNA	0.073	0.543
16	48600000	48700000	ENSBIXG00005016351	LRRC47	protein coding	0.073	0.543
16	48600000	48700000	ENSBIXG00005016324	SMIM1	protein coding	0.073	0.543
16	48600000	48700000	ENSBIXG00005016305	not available	protein coding	0.073	0.543
16	54600000	54700000	ENSBIXG00005008560	RC3H1	protein coding	0.091	0.43
16	54800000	54900000	ENSBIXG00005000199	RABGAP1L	protein coding	0.035	0.447
16	79400000	79500000	ENSBIXG00005019828	PPP1R12B	protein coding	0.05	0.423
16	79400000	79500000	ENSBIXG00005019795	RF00004	snRNA	0.05	0.423
19	27100000	27200000	ENSBIXG00005010360	not available	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005010357	MIS12	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005010333	DERL2	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005010287	DHX33	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005010261	NUP88	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005010218	RPAIN	protein coding	0.092	0.403
19	27100000	27200000	ENSBIXG00005001571	RABEP1	protein coding	0.092	0.403
19	46800000	46900000	ENSBIXG00005022992	CRHR1	protein coding	0.075	0.468
20	18500000	18600000	ENSBIXG00005013524	not available	protein coding	0.05	0.444
20	18500000	18600000	ENSBIXG00005013520	RF02160	misc RNA	0.05	0.444
20	18500000	18600000	ENSBIXG00005013509	RF02159	misc RNA	0.05	0.444
21	7800000	7900000	ENSBIXG00005023285	not available	lncRNA	0.09	0.401
21	7800000	7900000	ENSBIXG00005023277	not available	lncRNA	0.09	0.401
21	7800000	7900000	ENSBIXG00005023256	IGF1R	protein coding	0.09	0.401
21	49000000	49100000	ENSBIXG00005025684	SEC23A	protein coding	0.056	0.571
21	49000000	49100000	ENSBIXG00005004365	GEMIN2	protein coding	0.056	0.571
21	49000000	49100000	ENSBIXG00005025542	not available	protein coding	0.056	0.571
21	49000000	49100000	ENSBIXG00005004347	TRAPPC6B	protein coding	0.056	0.571
21	49000000	49100000	ENSBIXG00005025501	PNN	protein coding	0.056	0.571
23	14700000	14800000	ENSBIXG00005010033	KIF6	protein coding	0.094	0.488

Continued on next page

Table 6.5 – continued from previous page

Chr	Start	End	Ensembl ID	Name	biotype	Indicine mean proportion alternate allele	Taurine mean proportion alternate allele
23	43400000	43500000	ENSBIXG00005028562	not available	lncRNA	0.083	0.573
24	200000	300000	ENSBIXG00005023903	RF00001	rRNA	0.055	0.702
24	500000	600000	ENSBIXG00005006840	PARD6G	protein coding	0.064	0.544
24	600000	700000	ENSBIXG00005004059	ADNP2	protein coding	0.073	0.587
24	600000	700000	ENSBIXG00005023670	not available	protein coding	0.073	0.587
24	600000	700000	ENSBIXG00005023663	RBFA	protein coding	0.073	0.587
24	37200000	37300000	ENSBIXG00005013581	SMCHD1	protein coding	0.033	0.507
24	37200000	37300000	ENSBIXG00005013554	EMILIN2	protein coding	0.033	0.507
24	37200000	37300000	ENSBIXG00005013545	RF00026	snRNA	0.033	0.507
27	41300000	41400000	ENSBIXG00005026717	THRB	protein coding	0.017	0.574
27	41300000	41400000	ENSBIXG00005026549	NR1D2	protein coding	0.017	0.574
28	18800000	18900000	ENSBIXG00005020285	not available	protein coding	0.027	0.415
28	18800000	18900000	ENSBIXG00005020283	ADO	protein coding	0.027	0.415
28	18800000	18900000	ENSBIXG00005020273	EGR2	protein coding	0.027	0.415
29	19200000	19300000	ENSBIXG00005028250	ETS1	protein coding	0.1	0.458

Table 6.6: Annotation of SNP and INDEL variants.

Description	Angus	Brahman
number of animals	6	5
nonsynonymous SNV	53730	79170
stop gain	871	1253
stop loss	220	267
synonymous SNV	47843	96675
frameshift deletion	1350	1866
frameshift insertion	1120	1397
nonframeshift deletion	519	845
nonframeshift insertion	386	588
stop gain	61	101
stop loss	9	15

Short read data from either Angus or Brahman was mapped to the corresponding reference genomes. After GATK variant calling and filtering of variants, Annovar was used to annotate the variants identified. Note: SNV is single nucleotide variant.

Table 6.7: Breed-specific structural variant (SV) type and over/under-represented gene ontology for biological processes.

Angus-specific insertion SV				
PANTHER GO-Slim Biological Process	Over/Under-represented GO	Fold enrichment	Raw P-value	FDR
cellular response to stimulus (GO:0051716)	+	1.75	8.49E-06	1.52E-02
Angus-specific tandem contraction SV				
PANTHER GO-Slim Biological Process	Over/Under-represented GO	Fold enrichment	Raw P-value	FDR
synaptic vesicle endocytosis (GO:0048488)	+	24.92	3.75E-06	3.35E-03
synaptic vesicle cycle (GO:0099504)	+	19.29	1.14E-05	5.08E-03
organophosphate biosynthetic process (GO:0090407)	+	15.43	1.96E-04	3.18E-02
peptide metabolic process (GO:0006518)	+	13	6.40E-05	1.91E-02
regulation of cation transmembrane transport (GO:1904062)	+	12.46	7.71E-05	1.97E-02
regulation of ion transmembrane transport (GO:0034765)	+	12.46	7.71E-05	1.72E-02
regulation of ion transport (GO:0043269)	+	10.47	8.03E-06	4.78E-03
regulation of transport (GO:0051049)	+	7.9	4.43E-05	1.58E-02
membrane invagination (GO:0010324)	+	6.2	1.86E-04	3.70E-02
vesicle budding from membrane (GO:0006900)	+	6.2	1.86E-04	3.33E-02
regulation of localization (GO:0032879)	+	6.06	3.37E-06	6.03E-03
Brahman-specific insertion SV				
PANTHER GO-Slim Biological Process	Over/Under-represented GO	Fold enrichment	Raw P-value	FDR
release of sequestered calcium ion into cytosol (GO:0051209)	+	6.93	1.87E-04	2.78E-02
negative regulation of sequestering of calcium ion (GO:0051283)	+	6.67	2.29E-04	2.56E-02
phospholipid translocation (GO:0045332)	+	5.71	8.27E-05	2.11E-02
organophosphate biosynthetic process (GO:0090407)	+	5.59	5.72E-04	4.09E-02
lipid translocation (GO:0034204)	+	5.57	9.78E-05	2.19E-02
sequestering of calcium ion (GO:0051208)	+	5.57	9.78E-05	1.94E-02
positive regulation of cell migration (GO:0030335)	+	5.5	2.56E-04	2.54E-02
regulation of ion transport (GO:0043269)	+	3.71	2.28E-04	2.71E-02
lipid transport (GO:0006869)	+	3.32	3.42E-04	2.91E-02
negative regulation of cellular process (GO:0048523)	+	3.27	2.35E-04	2.47E-02
microtubule-based movement (GO:0007018)	+	3.19	7.74E-04	4.77E-02

Continued on next page

Table 6.7 – continued from previous page

phosphate-containing compound metabolic process (GO:0006796)	+	3.12	5.74E-04	3.94E-02
regulation of transport (GO:0051049)	+	3.04	7.32E-04	4.85E-02
lipid localization (GO:0010876)	+	3.04	7.32E-04	4.67E-02
regulation of membrane potential (GO:0042391)	+	2.96	1.56E-04	2.54E-02
organophosphate metabolic process (GO:0019637)	+	2.88	2.11E-04	2.90E-02
regulation of localization (GO:0032879)	+	2.74	2.50E-05	7.44E-03
membrane fusion (GO:0061025)	+	2.7	4.27E-04	3.18E-02
microtubule-based process (GO:0007017)	+	2.61	1.04E-04	1.86E-02
macromolecule localization (GO:0033036)	+	2.53	4.23E-04	3.29E-02
signal transduction (GO:0007165)	+	1.65	2.29E-07	1.36E-04
intracellular signal transduction (GO:0035556)	+	1.64	3.88E-04	3.15E-02
cellular response to stimulus (GO:0051716)	+	1.57	6.15E-07	2.20E-04
localization (GO:0051179)	+	1.41	2.21E-04	2.83E-02
cellular process (GO:0009987)	+	1.27	4.24E-07	1.90E-04
gene expression (GO:0010467)	-	0.62	3.32E-04	2.97E-02
sensory perception (GO:0007600)	-	0.35	2.75E-04	2.59E-02
sensory perception of chemical stimulus (GO:0007606)	-	0.05	4.59E-08	4.10E-05
detection of chemical stimulus involved in sensory perception (GO:0050907)	-	<0.01	1.12E-08	2.00E-05

Brahman-specific deletion SV

PANTHER GO-Slim Biological Process	Over/Under-represented GO	Fold enrichment	Raw P-value	FDR
organophosphate biosynthetic process (GO:0090407)	+	5.9	4.21E-04	4.18E-02
small molecule biosynthetic process (GO:0044283)	+	5.5	2.49E-04	3.43E-02
sequestering of calcium ion (GO:0051208)	+	5.22	3.38E-04	4.32E-02
calcium-mediated signaling (GO:0019722)	+	3.82	1.75E-04	2.84E-02
cellular calcium ion homeostasis (GO:0006874)	+	3.22	8.44E-06	5.03E-03
calcium ion homeostasis (GO:0055074)	+	3.21	9.16E-06	4.09E-03
negative regulation of cellular process (GO:0048523)	+	3.2	4.55E-04	4.28E-02
divalent inorganic cation homeostasis (GO:0072507)	+	2.98	1.64E-05	5.86E-03
ion homeostasis (GO:0050801)	+	2.43	5.87E-05	1.31E-02
inorganic ion homeostasis (GO:0098771)	+	2.37	1.10E-04	1.97E-02
chemical homeostasis (GO:0048878)	+	2.35	3.85E-05	9.82E-03
homeostatic process (GO:0042592)	+	2.05	3.56E-04	4.24E-02
regulation of biological quality (GO:0065008)	+	1.75	3.83E-04	4.02E-02

Continued on next page

Table 6.7 – continued from previous page

intracellular signal transduction (GO:0035556)	+	1.71	1.90E-04	2.83E-02
signal transduction (GO:0007165)	+	1.53	2.44E-05	7.27E-03
cellular response to stimulus (GO:0051716)	+	1.4	3.62E-04	4.05E-02
cellular process (GO:0009987)	+	1.22	7.56E-05	1.50E-02
sensory perception of chemical stimulus (GO:0007606)	-	0.1	1.30E-06	1.17E-03
detection of chemical stimulus involved in sensory perception (GO:0050907)	-	<0.01	2.33E-08	4.17E-05

The null hypothesis is all cattle genes and genes that overlapped with SV intervals are equally likely to be found in a particular GO- Slim Biological Process. We have used two-sided Fisher's Exact test and applied correction to the p-value using False Discovery Rate (FDR). The raw p-values are from Fisher's Exact test before FDR adjustment.

6.3 Supplementary Notes

6.3.1 Comparison of different Hi-C scaffolding programs

Three different scaffolders, 3D-DNA (Dudchenko et al., 2017), Proximo (Phase Genomics) and SALSA (Ghurye et al., 2019) were evaluated for building scaffolds using the following parameters.

For 3D-DNA, raw reads were aligned with juicer git commit d940e9e012a75822ff3f5a9ed7b3ecf08999df01 with the options `-z 'pwd'/reference/asm.fasta -y 'pwd'/reference/asm_MboI.txt -q phillippy.q -l phillippy.q -D software/juicer/ -d 'pwd' -p 'pwd'/reference/chr.sizes`. Scaffolding with 3D de novo assembly: version 170123 used the command `-m haploid -t 15000 -s 2 -c 30 asm.fasta merged_nodups.txt` for both haplotypes.

Phase Genomics' Proximo Hi-C genome scaffolding platform (git commit 145c01be162be85c060c567d576bb4786496c032) was used to create chromosome-scale scaffolds from the contig assembly as described in Bickhart et al (Bickhart et al., 2017). As in the LACHESIS method4, this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number of Sau3AI restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency and other statistical patterns in Hi-C data. Approximately 40,000 separate Proximo runs were performed to optimize the number of scaffolds and scaffold construction in order to make the scaffolds as concordant with the observed Hi-C data as possible.

Details for the SALSA2 run is given in the Methods section of the manuscript. We evaluated the performance of each scaffolder using a heuristic scoring method involving mapping genetic markers to each set of scaffolds. Each method was scored based on the inverse of the number of scaffolds required to cover the base chromosome (N), the difference in cumulative length of these scaffolds compared to the analog chromosome in ARS-UCD1.2 in megabases (L), and the number of contig order errors within each scaffold. The score function can be summarized by this equation:

$$\text{Score} = \frac{1}{N + i + b + l} \quad (6.1)$$

where “i” represents the number of contig inversions and “b” represents the number of continuity breaks detected in the genetic map based on their positions in ARS-UCD1.2. Scoring was performed on a per-chromosome basis with replacement of scaffolds that had previously mapped to other chromosomes. Using this method, SALSA2 gave most chromosomes with the highest scores and hence was the chosen to use for scaffolding.

6.3.2 Comparison of optical map based scaffolding approaches

6.3.2.1 *De novo* optical map assembly and haplotype resolution

Approximately 450 Gb of sequence, representing ~167x coverage, with molecule length >150 kb was aligned to the Brahman and Angus haplotig assemblies respectively to produce haplotype resolved scaffolds of each breed. Alignment of each molecule to the haplotigs of each breed was given a confidence score, which is the log of the alignment p-value. If the confidence score was 2 points higher for one breed haplotigs than the other, then the molecule was binned with the breed with the higher score. Molecule with alignments having almost equal score, defined as within 2 confidence score points of each other, were considered as homozygous and were randomly binned to one of the two breeds to keep the

coverage in homozygous and heterozygous regions uniform. In summary, 135 Gb of the molecules aligned to Angus only, 141 Gb of the molecules aligned to Brahman only, 110 Gb aligned to both Angus and Brahman and were binned evenly. About 64 Gb of molecules aligned to neither the Brahman nor the Angus haplotigs and were also binned randomly to one or the other breed.

For the optical map assembly, a pairwise comparison of all DNA molecules was used to create a layout overlap graph, which was then used to create the initial consensus genome maps. By realigning molecules to the genome maps and using only the best-matched molecules, the label positions on the genome maps were refined and used to validate the minimum assembly tiling path. Next, the software aligned molecules to genome maps and extended the maps based on the molecules aligning beyond the map ends. Overlapping genome maps were then merged. This process was repeated 5 times before a

final refinement step was applied to “finish” all genome maps.

To analyse the advantages of haplotype-resolved vs haplotype-unaware optical map construction to guide scaffolding, we generated scaffolds based on the conventional approach that used all Bionano molecules. With about 450 Gbp (molecules >150kbp) of molecules collected from the Brahman-Angus offspring, the de novo assembly was 3.37 Gbp with an N50 of 71.11 Mbp. This approach was not biased a priori by assigning parental haplotype and instead resolved haplotype in a de novo manner. The advantages of this approach were an increased sequence coverage as molecules were not split to each haplotype and there was no reliance on molecule alignment to haplotigs to bin them to each breed. However, as parental alleles were not separated prior to assembly, switching between parental alleles in the scaffolds is possible. Furthermore, as the scaffolding algorithm was unaware that the Angus contigs have no X sequences, the final scaffold length for the Angus assembly was longer than expected as the genome map used to guide scaffolding included the X chromosome. The Angus sequence that aligned to the X chromosome genome map likely belonged to the Y chromosome pseudoautosomal region, which is known to have high sequence identity with the X chromosome.

Table 6.8: Input dataset used to perform optical map-based assemblies using the haplotype-resolved versus the conventional haplotype unaware approach.

Description	Angus × Brahman Offspring	Angus-Selected	Brahman-Selected
Data collected (molecules > 150 kbp)	480 Gbp	222 Gbp	228 Gbp
Effective coverage of reference	126x	65x	65x
Assembly size	3.37 Gbp	2.79 Gbp	2.87 Gbp
Genome map N50	71.11 Mbp	33.97 Mbp	28.62 Mbp

6.3.2.2 Haplotype-resolved scaffold assembly

Haplotype-resolved scaffold assembly was performed using the contigs of Brahman and Angus separately, and the Bionano genome map was assembled using standard parameters in Bionano Access (Bionano Solve 3.2.1). For both the Brahman-selected and Angus-selected assemblies, about 98% of the sequences were incorporated into the final hybrid

assemblies with N50s of about 34 Mbp. The Angus-selected scaffold has a length of 2.53 Gbp while the Brahman-selected scaffold has a length of 2.64 Gbp. The N50s of the Angus-selected and Brahman-selected assemblies are limited by the breakage of the Brahman and Angus sequence assemblies in potentially random regions of the genome, which creates a bias in the molecule alignment step during molecule selection. During the scaffolding process, 29 and 36 discrepancies were identified in the Angus and Brahman scaffolds, respectively. These were most likely sequence chimeras, and breaks were introduced in the sequence contigs.

Table 6.9: Angus-selected Bionano assembly with Angus contigs.

Statistic	Original Bionano	Original sequence	Sequence used in scaffold	Scaffold	Scaffold + leftover unscaffolded sequence
Number of contigs	597	1747	397	217	1595
N50 (Mbp)	33.97	29.44	32.50	35.49	35.24
Total length (Mbp)	2790.07	2573.81	2512.48 (97.6%)	2526.00	2587.27

Table 6.10: Brahman-selected Bionano assembly with Brahman contigs.

Statistic	Original Bionano	Original sequence	Sequence used in scaffold	Scaffold	Scaffold + leftover unscaffolded sequence
Number of contigs	493	1585	421	154	1353
N50 (Mbp)	28.62	23.45	21.99	32.70	31.74
Total length (Mbp)	2867.64	2678.77	2632.15 (98.3%)	2644.13	2690.21

6.3.2.3 Conventional optical map scaffold assembly

Scaffolding between the Brahman-Angus offspring Bionano map with the Brahman sequence and the Angus sequence were also performed respectively (Bionano Solve 3.2.2). The assemblies have lengths of about 2.65 Gbp with N50 of 84 Mbp. Potential sequence errors (133 in the Angus and 65 in the Brahman) were detected and corrected while

running the Bionano scaffolding pipeline (Bionano Solve 3.2.2).

Table 6.11: Offspring scaffolding with Angus contigs.

Statistic	Original Bionano	Original sequence	Sequence used in scaffold	Scaffold	Scaffold + leftover unscaffolded sequence
Number of contigs	1026	1747	496	111	1414
N50 (Mbp)	71.11	29.44	29.44	84.07	84.07
Total length (Mbp)	3370.21	2573.81	2518.93 (97.9%)	2643.34	2693.03

Table 6.12: Offspring scaffolding with Brahman contigs.

Statistic	Original Bionano	Original sequence	Sequence used in scaffold	Scaffold	Scaffold + leftover unscaffolded sequence
Number of contigs	1026	1585	433	87	1282
N50 (Mbp)	71.11	23.45	21.82	84.31	84.31
Total length (Mbp)	3370.21	2678.77	2632.43 (98.3%)	2660.03	2705.93

6.3.3 Genome annotation of UOA_Brahman_1 using the NCBI annotation pipeline

The NCBI Eukaryotic Genome Annotation Pipeline was used to annotate genes, transcripts, proteins and other genomic features on the *Bos indicus* haplotype (GCF_00336969.5.1). The methodology for producing NCBI *Bos indicus* × *Bos taurus* Annotation Release 100 (AR 100) was as described for the UMD_CASPUR_WB_2.0 assembly (Burton et al., 2013). The evidence aligned to the genome and used for gene prediction was made up of transcript data from the same individual that provided the sample for the genomic sequence: 56,550 PacBio Iso-Seq reads from brain, liver, kidney, placenta, skeletal muscle, lung, and heart, and 1.7 billion RNA-Seq reads from brain, liver, lung, placenta and skeletal muscle. The other evidence used were: 8 billion RNA-Seq reads from 15 *Bos indicus* tissue samples, transcripts and proteins from *Bos taurus* (14,281 known RefSeq proteins, 19,584 GenBank proteins, 1,583,270 ESTs), and 52,350 human known RefSeq proteins.

The resulting annotation consists of 20,846 protein-coding genes, 16,398 of which have an ortholog to human. Only 135 protein-coding genes are missing a start or a stop codon and are marked as partial. Where necessary, the annotation pipeline introduced differences between the predicted models and the genomic sequence to compensate for frameshift-causing genomic insertions or deletions that are not supported by protein alignments. These “corrected” proteins are prefixed with ‘LOW QUALITY PROTEIN’ and should be considered as lower confidence. As an additional testament to the quality of the assembly, only 677 protein coding genes required such a “correction”.

Table 6.13: Comparisons of the number of corrected coding sequences in selected mammalian species.

Scientific name	Assembly name	Sequencing technology	Number of corrected CDS
<i>Bos taurus</i>	ARS-UCD1.2	PacBio; Illumina NextSeq500; Illumina HiSeq; Illumina GAII	622
<i>Bos indicus</i> x <i>Bos taurus</i>	UOA_Brahman_1	PacBio Sequel; PacBio RSII; Illumina NextSeq	677
<i>Capra hircus</i>	ARS1	PacBio	946
<i>Ovis aries</i> <i>musimon</i>	Oori1	Not listed	979
<i>Bos indicus</i>	<i>Bos_indicus_1.0</i>	SOLiD	1383
Bison Bison	Bison_UMD1.0	454; Illumina HiSeq	1448
<i>Ovis aries</i>	Oar_rambouillet_v1.0	HiSeq X Ten; PacBio RS II	1646
<i>Bubalus</i> <i>bubalis</i>	UOA_WB_1	PacBio	1943
<i>Bos mutus</i>	BosGru_v2.0	Illumina HiSeq; Illumina GA	1954
<i>Ovis aries</i>	Oar_v4.0	Illumina GAII; 454; PacBioRSII	4524

Although our Brahman reference (UOA_Brahman_1) was not polished with short reads using a pipeline such as Pilon used for ARS-UCD1.2, the number of corrected CDS is very similar. This suggests polishing with PacBio reads alone, without mixing reads representing different haplotypes, has resulted in a consensus assembly with less INDEL errors, which is a common problem that leads to correction of CDS necessary in the annotation step.

6.3.4 Genome annotation of UOA_Angus_1 and UOA_Brahman_1 using the Ensembl annotation pipeline

Four major classes of evidence were used to create a set of candidate transcripts: pooled high-quality Iso-Seq transcripts from seven tissues (brain, liver, lung, skeletal muscle, placenta, kidney, and heart) of the sequenced F1 hybrid fetus, short read RNA-seq data

from five tissues (brain, liver, lung, skeletal muscle, and placenta) of the F1 hybrid fetus, along with publicly available *Bos taurus taurus* and *Bos taurus indicus* data, human transcripts mapped from the GENCODE (Williams et al., 2017) gene set using pairwise whole genome alignment and finally vertebrate proteins with experimental evidence from UniProt (Frankish et al., 2019) (see below).

Table 6.14: Initial transcript models for each major input data type for Ensembl annotation.

Data type	Maternalinitial initial models	Paternalinitial initial models
Iso-Seq (7 pooled tissues)	165822	161968
Short read RNA-seq (32 tissues + 1 merged)	1167059	1150303
Human GENCODE mapping	53178	52242
UniProt known vertebrate proteins	542891	533051

The unfiltered transcript models came from the initial alignments of each datatype, and later in the annotation process low quality, fragmented and redundant models were removed to produce the finalized gene/transcript models.

Data from each locus was assessed to remove low quality models and then collapsed into a final gene model with an associated non-redundant transcript set. During the collapsing process, priority was given to transcript models based on the transcriptomic (Iso-Seq and RNA-Seq) data over models derived from homology. For protein coding genes, we also assessed the coverage of the open reading frame (ORF) in relation to known vertebrate proteins. The most complete model was chosen at each locus. In cases where the transcriptomic data appeared fragmented in comparison to the homology data, the homology data were included for completeness. Similarly, for regions where there were no transcriptomic data, we included models based on homology if there was a sufficiently good alignment.

Each gene was then classified as protein coding, long non-coding or pseudogene based on an analysis of the alignment information present at each locus. Genes with transcripts matching known proteins, which did not display multiple structural abnormalities, were classified as protein coding. If a gene matched a known protein but had several problems with the underlying structure (i.e., non-canonical splice sites, abnormally short introns, high level of repeat coverage, no evidence of expression), we classified it as a pseudogene.

Single exon genes were assessed for evidence of retrotransposition based on the presence of a multi-exon gene with a highly similar ORF elsewhere in the genome. Such single exon genes were classed as processed pseudogenes. If a gene fell into none of the previous categories, did not overlap a protein coding gene and had been constructed from transcriptomic data, it was considered as a potential lncRNA. The lncRNA set was filtered to remove transcripts that did not have at least two valid splice sites or cover 1000 bp.

In addition to the above, a small non-coding RNA annotation was produced. The miRNA genes were identified by running a BLAST (Consortium, 2019) of miRbase (Altschul et al., 1990) against the genome and then passing the results into RNAfold (Kozomara et al., 2019). Results were post filtered to remove poor quality alignments or alignments that were covered by repeats. For other small non-coding gene types, Rfam (Gruber et al., 2008) was used to scan against the genome and the results were passed into Infernal (Kalvari et al., 2018). Both the maternal and paternal gene sets are available as part of Ensembl release 97. More information on the annotation is available at http://asia.ensembl.org/info/genome/genebuild/2019_06_hybrid_cattle_gene_annotation.pdf.

6.3.5 Further assembly evaluation

The completeness of the genome from contig to chromosome-level assembly was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0.1 (Nawrocki and Eddy, 2013). The *mammalia_odb9* lineage-specific profile that contains 4,104 BUSCO gene groups was tested against the Brahman and Angus genome assemblies using the option `-m geno`. In addition to testing for completeness in gene space using BUSCO, the assembly base quality values (QV) and other assembly metrics including compression/expansion (CE) errors were calculated using BWA MEM (Simão et al., 2015), FRCBam (Li and Durbin, 2009) and Freebayes (Vezzi et al., 2012) as previously described (Garrison and Marth, 2012).

For consistency in the evaluation of some recently assembled mammalian genomes, we assessed the error rates using the same method described for the water buffalo (Koren et al., 2018) and goat (Bickhart et al., 2017) genomes. Briefly, short-insert Illumina WGS reads

from the parents of the sequenced hybrid F1 animal were aligned to the corresponding haplotype-resolved assemblies using BWA MEM (Simão et al., 2015). These short reads were not used in the assembly process and hence they served as an independent dataset to evaluate the genomes. We used the reference-free assembly validation software, FRCBam (Li and Durbin, 2009) to generate feature response curves on the Angus and Brahman assemblies in order to identify compression/expansion (CE) errors in the genomes. Putative erroneous bases in each assembly were also identified using FreeBayes (Vezzi et al., 2012). The results are tabulated in Supplementary Table 6.2. Commands used to generate all assembly quality assessment metrics can be found in the GitHub repository (<https://github.com/lloydlow/BrahmanAngusAssemblyScripts>).

6.3.6 Identification of selective sweep regions

To uncover genetic variants involved in indicine adaptative selection, we designed a strategy to identify selective sweeps using the haplotype-resolved Brahman genome. This method is analogous to the Cross Population Extended Haplotype Homozygosity (XP-EHH) (Low et al., 2019) in that genomic regions are searched for selected alleles that are approaching fixation in the Brahman population but remains polymorphic in six other taurine breeds populations. Only SNPs from the 38 individuals representing seven breeds from the USMARC Beef Diversity Panel version 2.9 (MBCDPv2.9) were considered in the selective sweep analysis.

6.3.6.1 The selection sweep method

If a genomic region has a high level of homozygosity within Brahman individuals, but the same region contains segregating polymorphic variants in taurine breeds (or visa versa), it can be considered as a candidate region for a selective sweep. An alternative explanation for such SNP patterns, which cannot be discounted on the available evidence, is genetic drift. Individuals chosen for selective sweep detection were sires with minimal pedigree relationships to ensure sampling of diverse alleles suitable for a population genetics study (Sabeti et al., 2007). The methodology for is similar to other recent studies to identifying selective sweep in cattle (Heaton et al., 2016, Koufariotis et al., 2018). The present study

had the advantage of using a haplotype-resolved Brahman genome to call SNPs rather than relying on the poorer resolution taurine-based reference, UMD3.1.1.

The proportion of alternate alleles in fixed size windows for the 38 animals representing seven breeds was calculated using the Brahman reference genome. To calculate the proportion of alternate alleles, the annotated genotype data was processed in R using custom scripts to ensure good quality data was included. Only genotypes with complete calls for all animals were retained. For Brahman genotypes, a further requirement of at least five mapped reads was imposed to ensure sufficient coverage to detect the presence of alternate alleles. A genotype labelled as 0/0 is homozygous and the same as the reference and hence there is no alternate allele. The maximum number of alternate alleles possible at a locus for an individual is two (e.g., a genotype call of 1/1).

Let

n = Number of individuals

c = Count of alternate allele at a particular position per individual

m = Number of SNPs in the window.

Proportion of alternate allele at position j is

$$\frac{\sum_{i=1}^n c_{ij}}{2n}, \quad (6.2)$$

where c_{11} refers to the count of alternate allele for the 1st individual at the 1st SNP in the window. Therefore, for each window, mean proportion of alternate allele (MP) is

$$\text{MP} = \frac{\sum_{j=1}^m \left(\frac{\sum_{i=1}^n c_{ij}}{2n} \right)}{m}. \quad (6.3)$$

Using this formula, indicine individuals were grouped together to calculate proportion mean of alternate allele, and the same process was repeated for the taurine individuals as a group. The figure below shows the proportion of mean alternate alleles in Brahman and the other six taurine breeds as histograms.

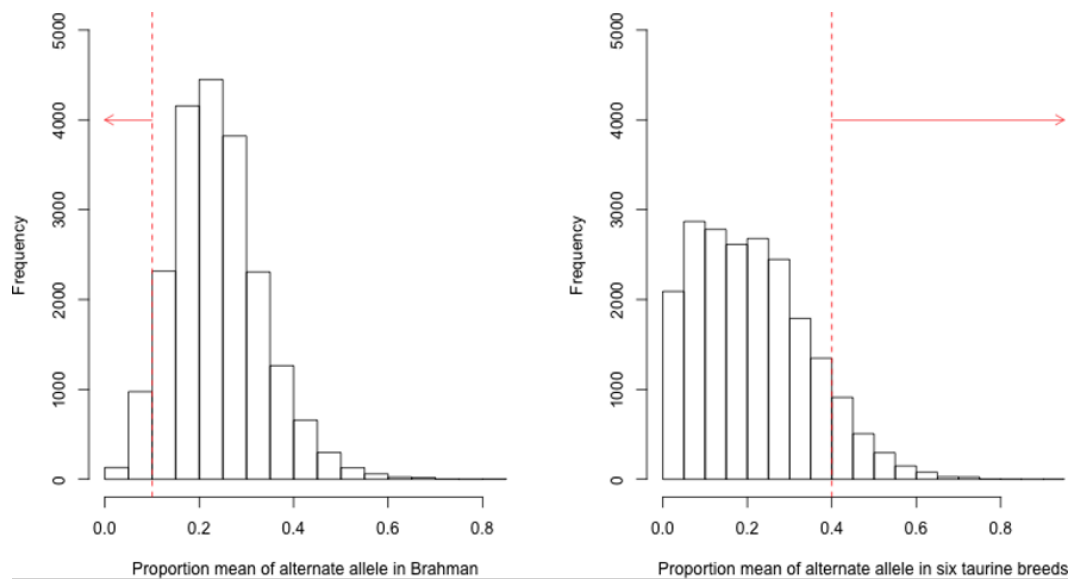


Figure 6.10: Histograms of the mean proportion alternate allele in Brahman and the other six taurine breeds.

The red line to the left in the Brahman plot indicates the bottom 5% percentile (i.e., extremely low polymorphism). The red line to the right in the taurine plot indicates the top 10% percentile with high polymorphic divergence relative to the Brahman.

Three fixed size windows (50 kb, 100 kb and 150 kb) were tested to check which would cover sufficient SNPs for the calculation of alternate alleles, and to determine the distance of consecutive SNPs. The table below shows the comparison of different window sizes for the dataset.

Table 6.15: Comparisons of different window sizes for SNP counts and consecutive SNP distances in each window.

Window size (kb)	Mean SNP count	Median SNP count	Mean consecutive SNP distance (bp)	Median consecutive SNP distance (bp)	Percentage of windows with at least 10SNPs
50	8.381	8.000	3341	3306	36%
100	16.39	16.00	5321	5028	81%
150	24.39	24.00	6122	5562	92%

The 50 kb window size was not appropriate for the selective sweeps analysis because many windows contained less than 10 SNPs. When the window size selected was 100 kb, 81% of all windows contained at least 10 SNPs. Additionally, the distance between consecutive SNPs was ~ 800 bp shorter, on average, when compared to the 150 kb window size. In other words, there was higher density of SNPs when 100 kb windows were chosen. Regardless of whether the 100 kb or the 150 kb window size was chosen, the final gene

list in the selection intervals overlapped substantially, although more candidate genes were detected using 100 kb windows. Only windows with at least 10 SNPs were considered for calculation of mean proportion of alternate alleles.

6.3.6.2 Pathway analysis

Candidate selective sweep genes in the Brahman genome were analyzed for potential genes overrepresentation in biological pathway using PANTHER 14.1 (Thomas et al., 2003). As there are no *Bos taurus indicus* specific biological pathways annotated, the candidate Brahman genes were mapped to the corresponding Hereford ARS-UCD1.2 Ensembl annotation release96 prior to running PANTHER. The selective sweep intervals were also searched against cattle quantitative trait loci (QTL) from Animal QTL database (Hu et al., 2013).

6.3.6.3 Interpretation of selective sweep results

We compared the SNP patterns of 5 Brahman with 33 individuals from six taurine breeds to identify signatures of selective sweeps (Figure 6.11a). We searched 100 kb windows spanning the whole genome for those where there was high level of homozygosity within Brahman individuals but with more segregating polymorphic variants in taurine breeds. This identified a total of 128 genes in 60 selective sweep intervals. Among these candidate selected genes, 80% were protein-coding, 1 was a pseudogene and the remainder were RNA-based genes (Supplementary Table 6.5). No biological pathways were found to be significantly over-represented among the positively selected protein-coding genes. The heat shock protein HSPA4, a member of the Hsp70 family, was amongst genes identified as under selection (Figure 6.11b). This region was also found in a search for selective sweeps in African cattle (Kim et al., 2017), which we would not expect by chance. We also identified *DNAJC13*, a member of a gene family known to act as co-chaperones of heat-shock proteins, in another selective sweep region.

A total of 231 unique QTL covering all six major QTL types listed in the cattle QTL database overlapped with positively selected genomic regions. The major QTL types were reproduction (26%) followed by exterior phenotype (22.9%) and milk (22.1%) traits

(Figure 6.11c). Ten of 60 selective sweep intervals did not overlap with any of the currently identified QTL.

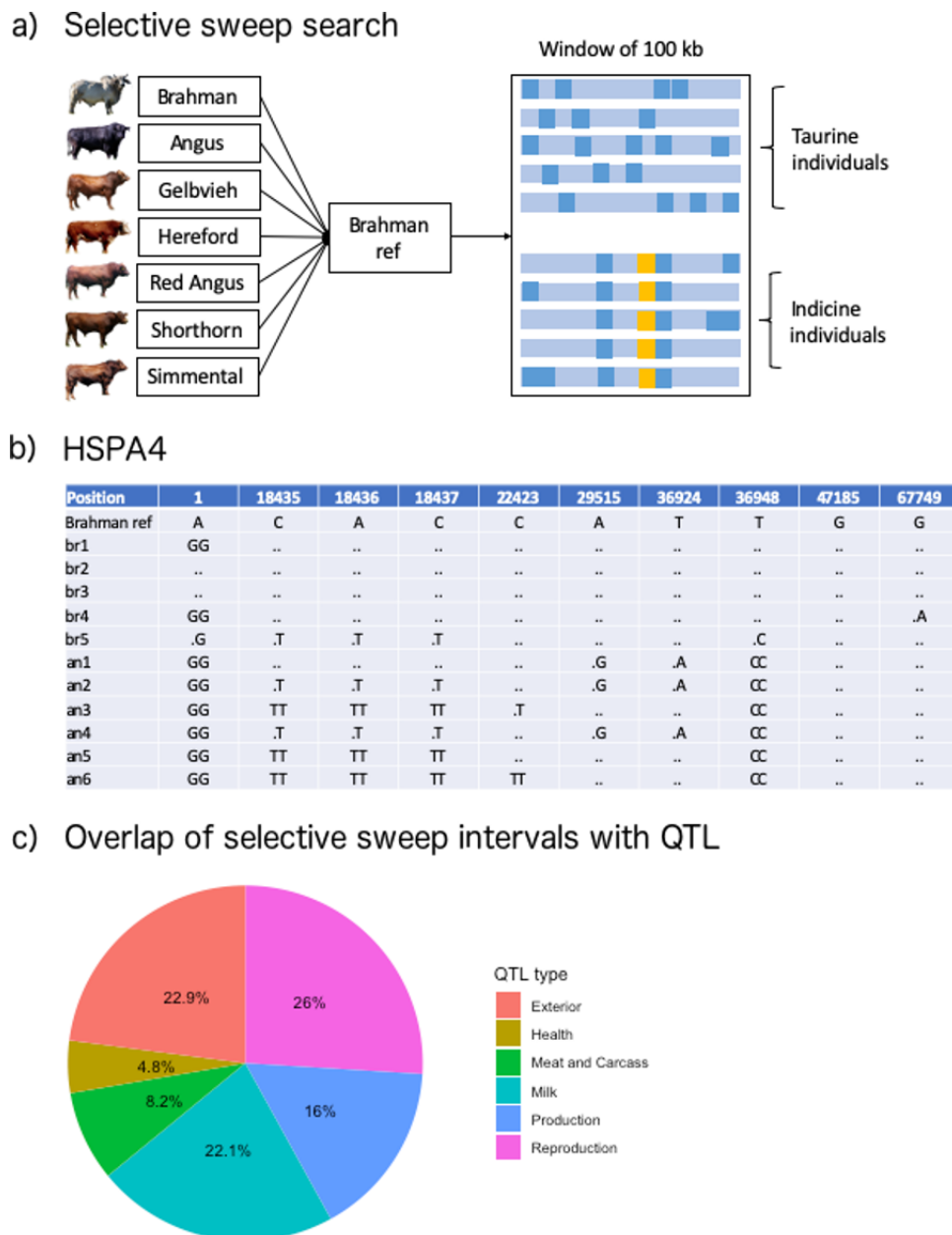


Figure 6.11: Selective sweep analysis in Brahman.

a) An overview of the strategy used to identify selective sweep regions by aligning and calling SNPs in 100-kb windows on the Brahman reference. The variant under selection is highlighted in yellow. b) An example of genotype calls for *HSPA4*, a gene residing in one of the selective sweep intervals, using partial results from Angus and Brahman individuals. Genotype shown as “.” means it follows the Brahman reference, which is haploid, and “an” denotes Angus whereas “br” denotes Brahman. The position indicates adjusted position starting from 66,130,388 bp in UOA_Brahman_1, chromosome 7. c) Overlap of selective sweep intervals with cattle QTL categorized by types. Only unique QTL IDs were used. The significant differences in phenotype, energy metabolism and adaptation to heat stress

of indicine cattle have been linked to the thyroid hormone axis (Cowley et al., 1971, Obeidat et al., 2002, Façanha et al., 2019). One of the selective sweep regions in Brahman contains thyroid hormone receptor β (THRB), a ligand-activated pleiotropic transcription factor that modulates the expression of a large number of genes (Ortiga-Carvalho et al., 2014). Thyroid hormones are intrinsically connected to the growth hormone - insulin-like growth factor axis (*GH-IGF*) (Forhead and Fowden, 2014). Insulin-like growth factor 1 receptor (IGF1R) was found in another selective sweep region. Polymorphisms in *IGF1R* have been associated with age of puberty in Brahman cattle (Fortes et al., 2013). In comparison with taurine cattle Brahman tend to reach puberty late, which may have been under positive selection as a consequence of adaption to harsh tropical environments, ensuring that cows are more mature and robust at the time of first calving.

Quantitative trait loci for reproduction featured prominently in the comparison of Brahman selective sweep regions with known cattle QTL. Amongst the reproduction traits, QTL related to calf size and calving ease were overrepresented. Brahman cows deliver a small calf that is less likely to result in dystocia and still birth (Comerford et al., 1987), which is one major benefit of the introgression of indicine genetics into more productive taurine breeds. Selective sweep regions thus provide candidate genes for maternal control of birth weight.

6.3.7 Further Iso-Seq analysis

After exclusion of SNPs with less than 40-fold Iso-Seq read coverage and those in non-transcribed regions, IsoPhase identified 5806 genes with 52,270 phased transcripts.

6.3.7.1 SNP calls missed by IsoPhase are either in homopolymer regions or have low coverage

There are 8,093 (substitution) SNP calls that are missed by IsoPhase but called jointly by RNA-Seq and genomic data. 5589 (69%) of the missed calls are either within or adjacent to a homopolymer (HP) region. IsoPhase defines a HP region as a stretch of 4 consecutive identical bases and will not call a SNP if it is inside a HP or immediately adjacent to a HP region.

Of the remaining 1813 missed calls (22%), have effective base coverage less than 40 in the Iso-Seq data. Effective base coverage is different from full-length Iso-Seq read coverage because after alternative splicing, certain exons, introns, or UTRs may have lower coverage despite meeting the initial 40 fold threshold requirement to run IsoPhase.

Of the missed PacBio calls, 91% are either in or adjacent to HP regions, or have insufficient base coverage. Manual inspection of the remaining 9% of the missed calls suggest that either they were also adjacent (but not immediately) to HP regions, or had low coverage resulting in insignificant P-values to pass the IsoPhase SNP call threshold.

6.3.7.2 IsoPhase-unique SNP calls are dominantly A to G calls, which suggests RNA editing

There are 2830 SNP calls unique to IsoPhase, 1776 SNP calls unique to RNA-seq, and 2651 SNP calls unique to genome. We tally the unique calls by Reference to Alternative SNP call in the figure below using the transcribed orientation as the sense strand and show that A to G is the most dominant unique call in IsoPhase and RNA-Seq data. It is not clear why PacBio Iso-Seq is the only platform to call certain SNPs.

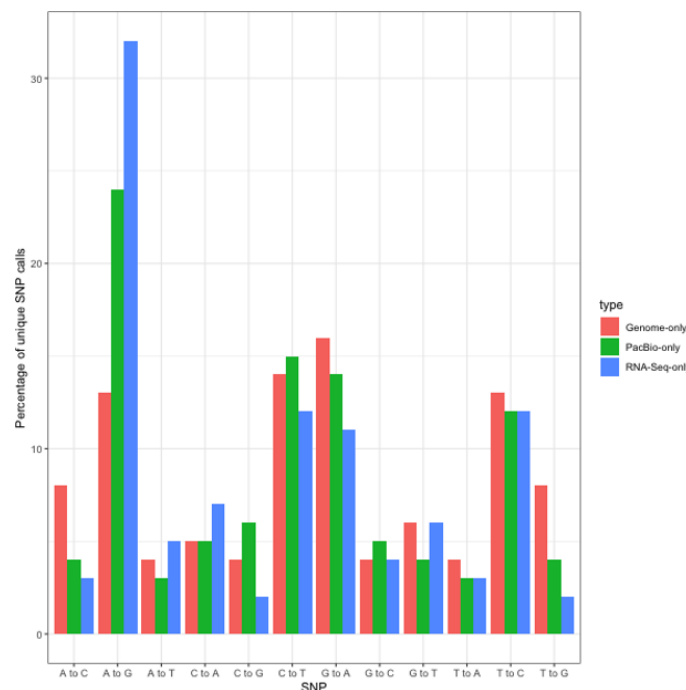


Figure 6.12: Percentages of SNP type across those unique in each of PacBio Iso-Seq, RNA-Seq and genome WGS datasets.

6.3.8 Rarefaction analysis of covered genes and transcripts

The subsampling of full-length non-concatamer (FLNC) reads at the level of genes and transcripts showed that brain, kidney and heart datasets were reaching a plateau, suggesting sequencing depth was adequate to discover the majority of transcripts for the annotation process. To ensure saturation of transcripts, additional data was produced for the brain sample, which was run with seven SMRT cells. By extrapolating from the rarefaction curve, 10 SMRT cells per tissue to gather ~1 million FLNC should ensure saturation of transcripts at this developmental stage.

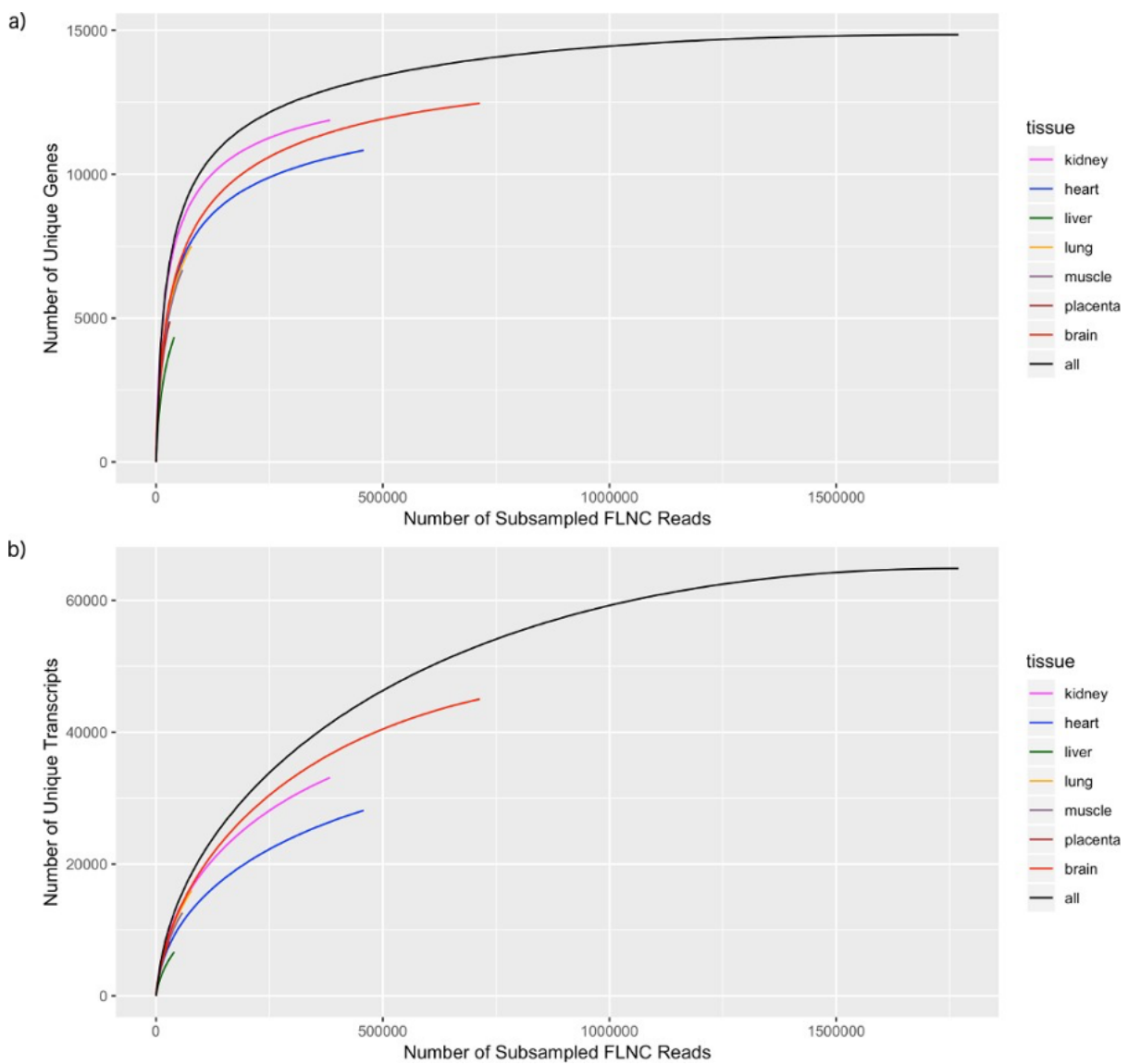


Figure 6.13: Rarefaction analysis of seven tissues in F1 animal at the level of a) gene, b) transcript.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.
- Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., Sullivan, S.T., et al., 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature genetics*, 49(4), pp.643–650.
- Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J., 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology*, 31(12), pp.1119–1125.
- Comerford, J., Bertrand, J., Benyshek, L., and Johnson, M., 1987. Reproductive rates, birth weight, calving ease and 24-h calf survival in a four-breed diallel among simmental, limousin, polled hereford and brahman beef cattle. *Journal of animal science*, 64(1), pp.65–76.
- Consortium, U., 2019. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1), pp.D506–D515.
- Cowley, J., Gutierrez, J., Warnick, A., Hentges Jr, J., and Feaster, J., 1971. Comparison of thyroid hormone levels in hereford and brahman cattle. *Journal of animal science*, 32(5), pp.981–983.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al., 2017. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, 356(6333), pp.92–95.
- Façanha, D.A.E., Ferreira, J.B., Leite, J.H.G.M., Sousa, J.E.R. de, Guilhermino, M.M., Costa, W.P., Asensio, L.A.B., Vasconcelos, A.M. de, and Silveira, R.M.F., 2019. The dynamic adaptation of brazilian brahman bulls. *Journal of thermal biology*, 81, pp.128–136.

- Forhead, A.J. and Fowden, A.L., 2014. Thyroid hormones in fetal growth and prepartum maturation. *Journal of endocrinology*, 221(3), R87–R103.
- Fortes, M.R., Li, Y., Collis, E., Zhang, Y., and Hawken, R.J., 2013. The igf 1 pathway genes and their association with age of puberty in cattle. *Animal genetics*, 44(1), pp.91–95.
- Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al., 2019. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1), pp.D766–D773.
- Garrison, E. and Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *Arxiv preprint arxiv:1207.3907*.
- Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., and Koren, S., 2019. Integrating hi-c links with assembly graphs for chromosome-scale assembly. *Plos computational biology*, 15(8), e1007273.
- Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L., 2008. The vienna rna websuite. *Nucleic acids research*, 36(suppl_2), W70–W74.
- Heaton, M., Smith, T., Carnahan, J., Basnayake, V., Qiu, J., Simpson, B., and Kalbfleisch, T., 2016. P6026 using diverse us beef cattle genomes to identify missense mutations in *epas1*, a gene associated with high-altitude pulmonary hypertension. *Journal of animal science*, 94(suppl_4), pp.161–162.
- Hu, Z.-L., Park, C.A., Wu, X.-L., and Reecy, J.M., 2013. Animal qtldb: an improved database tool for livestock animal qtl/association data dissemination in the post-genome era. *Nucleic acids research*, 41(D1), pp.D871–D879.
- Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I., 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding rna families. *Nucleic acids research*, 46(D1), pp.D335–D342.

- Kim, J., Hanotte, O., Mwai, O.A., Dessie, T., Bashir, S., Diallo, B., Agaba, M., Kim, K., Kwak, W., Sung, S., et al., 2017. The genome landscape of indigenous african cattle. *Genome biology*, 18(1), pp.1–14.
- Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hienleder, S., Williams, J.L., Smith, T.P., and Phillippy, A.M., 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, 36(12), pp.1174–1182.
- Koufariotis, L., Hayes, B., Kelly, M., Burns, B., Lyons, R., Stothard, P., Chamberlain, A., and Moore, S., 2018. Sequencing the mosaic genome of brahman cattle identifies historic and recent introgression including polled. *Scientific reports*, 8(1), pp.1–12.
- Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S., 2019. Mirbase: from microRNA sequences to function. *Nucleic acids research*, 47(D1), pp.D155–D162.
- Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.
- Low, W.Y., Tearle, R., Bickhart, D.M., Rosen, B.D., Kingan, S.B., Swale, T., Thibaud-Nissen, F., Murphy, T.D., Young, R., Lefevre, L., et al., 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature communications*, 10(1), pp.1–11.
- Nawrocki, E.P. and Eddy, S.R., 2013. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22), pp.2933–2935.
- Obeidat, B., Thomas, M., Hallford, D., Keisler, D., Petersen, M., Bryant, W., Garcia, M., Narro, L., and Lopez, R., 2002. Metabolic characteristics of multiparous angus and brahman cows grazing in the chihuahuan desert. *Journal of animal science*, 80(9), pp.2223–2233.
- Ortiga-Carvalho, T.M., Sidhaye, A.R., and Wondisford, F.E., 2014. Thyroid hormone receptors and resistance to thyroid hormone disorders. *Nature reviews endocrinology*, 10(10), p.582.

- Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al., 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164), pp.913–918.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M., 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), pp.3210–3212.
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A., 2003. Panther: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9), pp.2129–2141.
- Vezi, F., Narzisi, G., and Mishra, B., 2012. Reevaluating assembly evaluations with feature response curves: gage and assemblathons. *Plos one*, 7(12), e52210.
- Williams, J.L., Iamartino, D., Pruitt, K.D., Sonstegard, T., Smith, T.P., Low, W.Y., Biagini, T., Bomba, L., Capomaccio, S., Castiglioni, B., et al., 2017. Genome assembly and transcriptome resource for river buffalo, *bubalus bubalis* (2 n= 50). *Gigascience*, 6(10), gix088.